Retrieving documents based on gene name variations: MedIER at TREC 2017 Precision Medicine Track

Tong Yin
School of Information
University of Michigan
tongyin@umich.edu

Danny T. Y. Wu
Department of Biomedical Informatics
University of Cincinnati
wutz@ucmail.uc.edu

V. G. Vinod Vydiswaran
Department of Learning Health Sciences
University of Michigan
vgvinodv@umich.edu

***Abstract*—The TREC 2017 Precision Medicine Track focused on finding relevant medical documents – scientific abstracts and clinical trials – for cancer patient cases based on specific genetic variation and demographic information. We focused on the genetic variations mentioned in the query and explored ways to modify the search query and the retrieval ranking using this information. Further, we explored filtering retrieved results based on demographic information matching for clinical trials. The results show little improvements of the approaches over baseline runs, and suggest need for additional exploration.**

***Keywords*—gene name variation, query modification**

I. Introduction

In its previous incarnation, the Precision Medicine ran as the Clinical Decision Support track from 2014 to 2016. The Clinical Decision Support track focused on retrieving biomedical articles relevant to answering generic clinical questions about diagnosis, treatment, or test procedure for a given query topic. This year's task was primarily different in two areas: first, the query topics were focused on cancer diagnoses with specific genetic variations, and second, demographic information was available to further filter most relevant documents for the given query. The track included two tasks corresponding to retrieving (a) scientific abstracts and (b) clinical trials relevant to specific cancer cases.

Our participation in TREC 2017 Precision Medicine track was a collaborative endeavor between the University of Michigan and the University of Cincinnati. We participated in both tasks, and our approach focused on two research directions: (a) enhancing search queries with genetic variation information, and (b) using demographic information to select relevant clinical trials.

II. Related work

Since the Precision Medicine track grew out of the Clinical Decision Support track, the previous teams that participated in those track tasks were of immediate interest to us. In previous tracks, systems were expected to retrieve biomedical articles that were relevant to answering questions about the diagnoses, treatment plans, or test procedures related to clinical case reports. The case reports included information about a patient reported complaints, test results, and observations from the first few hours of patients' visit to the hospital.

We conducted a survey of participating systems for TREC 2014 [1], TREC 2015 [2], and TREC 2016 Clinical Decision Support tracks [3]. Our analysis of the challenge reports showed that most participating teams used pseudo-relevance feedback to improve the ranking results. Team MERCK-KGAA [4] used pseudo-relevance feedback to expand initial query by adding words of the titles of the top k biomedical articles. MayoNLP [5] team used pseudo-relevance feedback model to utilize co-occurring MeSH heading terms to expand the query topics. in addition, the queries were expanded using (a) document's keyword meta-information field, (b) high-ranking TF-IDF terms from the title, abstract, and the full article when available, and (c) MeSH or UMLS concepts from the title and abstract.

Other teams use negation-aware ranking, but it was not universally beneficial. The ETH Zurich team used a BM25 variant that could detect natural language negations by converting "no diabetes" to "[nx]diabetes." [6] They combined both the original terms and the negation terms, and learned the weight function during training. The team performance was above the baseline performance. Team SCIAICL also considered use of negated concepts and achieved close to baseline performance. [7]

Another key feature deployed by many teams was the use of word embedding models. Well-trained models were found to generate useful features by the top performing teams, while other teams suffered from setting less suitable parameters and reported poorer results. Team MERCKKGAA used word embedding to calculate document similarity between document centroids of topics and articles [4]. They found that such an approach contributed to a significant improvement in overall ranking. Team CBUN constructed semantic word vectors using the medical terms on word embedding. [8]. ETH Zurich [6] used the modified version of word2vec to expand to k neighbors and maximizing the cosine similarity with the given query. Although this was an interesting direction, we could not use this approach for TREC 2017 tasks because the disease name and gene variants need to match exactly.

Learning to rank was also a popular choice among the TREC 2016 clinical decision support track participants. Team SCIAICL used it to determine the weight for symptom queries [7], Team MERCKKGAA used learning to rank with gradient boosting to maximize NDCG. [4] Both teams were

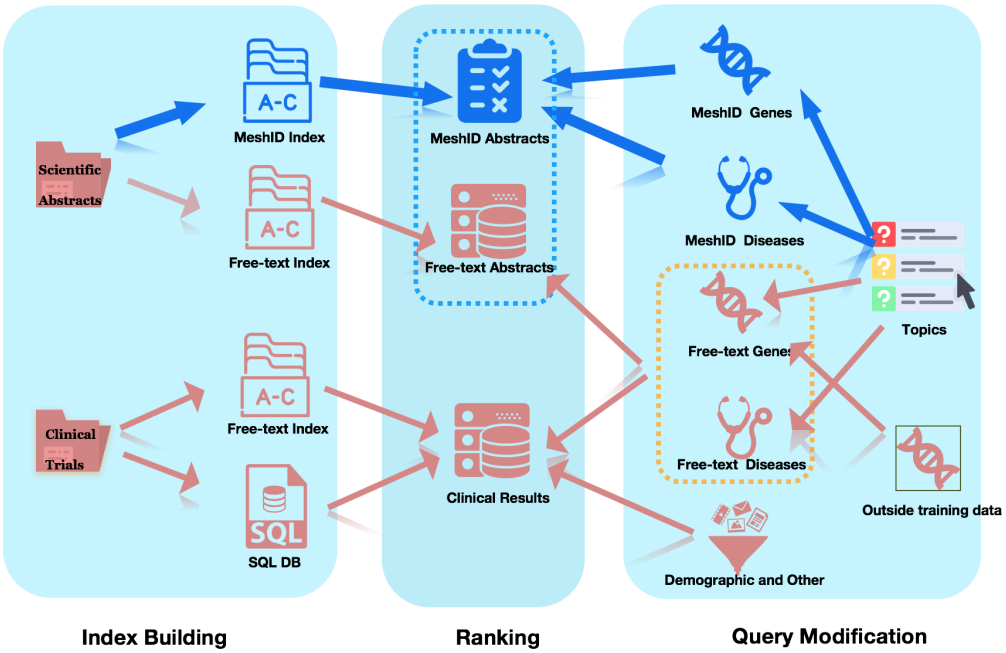

Fig. 1: System architecture. Blue color represents components using MeSH identifiers, while red color represents free text.

able to improve their results above the baseline performance. However, because there was no specific training dataset for this year's tasks, we could not use this approach.

Finally, some teams made use of Wikipedia or results from Google search directly. Team CBUN found the corresponding symptoms for each diseases using Wikipedia and created the clinical causal relationships [8]. Team PRNACL used the Wikipedia clinical medicine category pages to build a directed knowledge graph [9], with symptoms included as leaf nodes. Team CSIRO [10] used Wikipedia to expand names of diseases. Team iRiS built a Wikipedia index to predict the patient diagnosis [11]. Some of these approaches seemed relevant in this year's task, especially with respect to handling disease names.

III. Approach

Our approach focused on two research directions: (a) enhancing search queries with genetic variation information, and (b) using demographic information to select relevant clinical trials. For this track, all participating teams were provided two datasets – a collection of scientific abstracts of peer-reviewed biomedical research abstracts from PubMed, and a collection of clinical trials from ClinicalTrials.gov. The goal of the track was to identify the most relevant scientific abstracts and clinical trials pertaining to a given query, where a query is specified as a type of cancer, a specific gene variation, and some patient demographic information. The demographic information is especially important in filtering out clinical trials for which the patient may not be eligible.

A schematic diagram of our system architecture is shown in Fig. 1. Our system consisted of three components that are fairly common in any retrieval system, viz. (a) pre-processing and indexing the corpora, (b) query modification, and (c) retrieval and ranking. These components are explained in detail in the following sections.

A. Pre-processing and building index

The scientific abstracts corpus mainly comes from the January 2017 snapshot of PubMed abstracts, along with the abstracts obtained from AACR and ASCO proceedings. The overall corpus consists of 27.8 billion articles. In addition to the title and abstracts of each scientific paper in plain text form, the dataset also included some metadata about the journal publication dates, history, author, and MeSH identifiers of medical terms that appeared in that document. We parsed all available data in the XML version of the corpus to build two separate indexes: one index over the free text of article titles and abstracts, which another on the MeSH identifiers of medical terms noted for each article.

The clinical trials corpus consists of 176,000 trials and includes title, summary, and a detailed description. In addition, they also specify the eligibility criteria on age and gender for cohorts included in the trial. To specifically match and filter trials based on age and gender, we built an SQL database to store the age and gender specific eligibility criteria for every trial. The SQL database was later used in the retrieval and ranking phase to filter out non-eligible trials.

B. Query modification

A set of 30 query topics were released as part of the TREC task, each with a set of four relevant factors: type of cancer, relevant genetic variation, demographic criteria, and other pertinent information. Since the scientific abstracts have

no specific demographic information, we use the demographic information only for the clinical trial task.

1) Identifying candidate synonyms for gene names:: One of our primarily goals in participating for this year's challenge was to understand how gene variations are expressed in relevant documents and to enrich the keyword queries with information about the genetic variants. In preparing for the task, we collected information about gene names and synonyms from the National Center for Biotechnology Information (NCBI). A dataset of gene names and PubMed articles was selected from the NCBI gene resource[1]. Every gene is assigned a unique identifier, and are then listed as metadata in scientific articles on PubMed if the article mentions the specific gene. Hence, this resource provides a many-to-many relationship between genes and peer-reviewed research articles. For example, the PubMed ID 9873079 has four corresponding gene ID 1246502 "leuA", 1246503 "leuB", 1246504 "leuC", and 1246505 "leuD". The specific geneID 1246502 also has another related PubMed article with ID 9812361. The gene name and synonym resource generated our candidate list of synonyms.

2) Pruning ambiguous synonyms:: In addition to gene names, the NCBI dataset also includes the most relevant PubMed articles corresponding to the genes. We used these documents in a relevance feedback setup to prune out candidate synonyms that either do not contribute to finding relevant articles or are ambiguous and lead to retrieving many non-relevant articles. We begin with all gene synonyms as a query and retrieve relevant articles from the free-text scientific abstract index. Based on the NCBI dataset, we calculate the baseline precision of the retrieved results. Then, we remove the gene synonyms one-by-one and compute the precision of the retrieved results at every step. If the precision *increases* above the baseline precision, the synonym will be pruned. For example, gene "cdkn2a" has geneID of 1029 in the NCBI dataset. The dataset has 2,031 PubMed articles related to geneID 1029. According to the gene name database, there are sixteen synonyms (aliases) of "cdkn2a", including "arf", "mlm", "p14", "p16", etc. Using all sixteen synonyms, we identified 80 relevant documents. It should be noted that the index is built over just the title and abstract text, while the NCBI dataset has access to the full paper to search for gene name variants. This contributes for the relatively low precision (80 instead of 2,031). If we remove "p16" from the synonym query, the number of relevant document retrieved increases to 189. This indicates that "p16" was too ambiguous, since including it in the query results in many non-relevant articles in the retrieved results. In contrast, if we only remove "p14", the number of relevant documents in the retrieved set drops to 58, implying that "p14" is necessary to retrieve at least 22 additional relevant documents. So, we keep "p14" in the gene name expansion set. This procedure is followed for every gene synonym over all the gene names in the query. Although time-consuming, this one-time pre-processing helps in selecting the most representative synonyms for gene names.

3) Expanding free-text queries:: Finally, we combined the gene names and filtered synonyms with the cancer type keywords. We expanded the names of the cancer diagnoses with common synonyms. For example, "skin cancer" and "melanoma"; "stomach cancer" and "gastric cancer", etc. These alternate names were gathered from the MeSH disease tree.

4) Query using MeSH identifiers: Our past experience in previous years' TREC tasks demonstrated that adding MeSH identifiers could significantly improve the performance of retrieval result [12]. Hence, we create a new query based on MeSH terms and used it to query the MeSH identifier index (the second free-text index) and retrieve additional documents.

C. Retrieval Process

As the third and final component of our system, we use the Galago toolkit[2] for our retrieval step. Galago deploys an inference network based retrieval model. We build two sets of queries for gene names and diseases. One is MeSH identifiers-based query that is run against the MeSH identifier index; while the second is a text query that runs against the free-text scientific abstracts index. The two retrieved results are merged to get a final ranked list. The documents which appear in both MeSH identifiers-based retrieval and free-text based retrieval are ranked at the top of the re-ranked result set in the order in which they appeared in the free-text query result. These documents are followed by the documents that only appear in MeSH identifiers result; followed by documents that appear only in the free-text retrieval results. The ranking order is based on the original normalized ranking scores in the corresponding ranked list.

For the clinical trials dataset, a similar approach was followed. However, since there was no MeSH identifier information available for the clinical trials, we ran the queries only against the free text index. On the other hand, the demographic information such as age and gender of eligible patients was provided in the clinical trail description. These demographic factors are matched against the characteristics of the given patient case (query topic). If a retrieved clinical trial specifically articulates gender or age criteria, and these criteria do not match with the given query, then the clinical trial is removed from the retrieved results. Trials with no eligibility criteria or the ones that match the given query topic are returned in the original retrieved order.

IV. Submitted runs and Results

We submitted four runs for both the scientific abstract and clinical trial tasks. Table I summarizes all the runs submitted by our team. The runs consist of two baseline runs and two runs varying the fusion algorithms deployed to combine and re-rank the retrieved documents.

A. Scientific Abstracts

For the scientific abstracts task, the baseline free-text query and baseline MeSH identifier query were submitted as MedIER_sa2 and MedIER_sa4, respectively. The re-ranked list obtained by running pseudo-relevance feedback on free text query was submitted as the MedIER_sa3 run. The final run, MedIER_sa1, was generated by merging the ranking results from the free-text query and the MeSH identifier

[1]Can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/

[2]https://www.lemurproject.org/galago.php

Task	RunID	MeSH ID Disease	MeSH ID Gene	Disease	Gene	Demographic	Other	Pseudo-relevance Feedback	Expanded Retrial
Scientific Abstracts	MedIER_sa1	Yes	Yes	Yes	Yes	No	Yes	No	No
	MedIER_sa2	Yes	Yes	No	No	No	No	No	No
	MedIER_sa3	Yes	Yes	No	No	No	Yes	Yes	No
	MedIER_sa4	No	No	Yes	Yes	No	No	No	No
Clinical Trials	MedIER_ct1	No	No	Yes	Yes	Yes	Yes	No	No
	MedIER_ct2	No	No	Yes	Yes	Yes	Yes	Yes	No
	MedIER_ct3	No	No	Yes	Yes	Yes	Yes	No	Yes
	MedIER_ct4	No	No	Yes	Yes	Yes	Yes	Yes	Yes

TABLE I: Summary of submission runs

Runs	infNDCG	P@10	R-prec
MedIER_sa1	0.2036	**0.3300**	0.1177
MedIER_sa2	**0.2103**	0.2967	**0.1326**
MedIER_sa3	0.1986	0.2733	0.1143
MedIER_sa4	0.1774	0.2800	0.1061
Best	0.5856	0.8600	0.3950
Median	0.2766	0.3733	0.1761

TABLE II: Scientific Abstracts Results. All four submitted runs are compared against the best and averaged median performance on three measures: inferred NDCG, precision at 10, and precision at recall of 100%.

query. The ranking function used for merging was: $(3000 - Ranking\_sa2) + (3000 - Ranking\_sa4)$.

The official results of our scientific abstract runs along with benchmarking runs from other participants is shown in Table II. The results show three measures: inferred NDCG (infNDCG), precision at 10 (P@10), and precision at recall of 100% (R-prec).

The run generated using MeSH identifiers for gene and disease (MedIER_sa2) was the best one on infNDCG and R-prec measures. This is consistent with our prior experiments on the importance of MeSH terms in finding relevant results. The combination of MeSH terms and free-text query (MedIER_sa1) performs the best on the P@10 measure. This is also consistent with our expectation that relevant documents that appear in both MeSH-based queries and free-text queries are truly relevant. The results indicate that additional information could lead to even better performance among the top candidate documents, while introducing some noise. When we compare the runs that enabled pseudo-relevance feedback against the baseline free-text query run, we notice that the pseudo-relevance feedback based run performed better on the infNDCG and R-prec measures, but the original free-text query performed better on the P@10 measure.

B. Clinical Trials

We submitted four runs for the clinical trials task. The first run, MedIER_ct1, is the generated using free-text queries, followed by the filtering step on demographics characteristics of age and gender. The second run, MedIER_ct2, is generated using similar approach followed by pseudo-relevance feedback to augment the results. Since the first two runs yielded fewer results, we expanded the number of retrieval documents to 3000 using pseudo-relevance feedback, followed by the demographic criteria filter to generate the other two runs, MedIER_ct3 and MedIER_ct4.

The official results of our clinical trials runs along with benchmarking runs from other participants is shown in Table III. The results show precision measures at three levels: at 5, 10, and 15 retrieved documents.

Run ID	P@5	P@10	P@15
MedIER_ct1	0.1724	0.1621	0.1379
MedIER_ct2	**0.1862**	**0.1724**	**0.1563**
MedIER_ct3	0.1724	0.1621	0.1379
MedIER_ct4	**0.1862**	**0.1724**	**0.1563**
Best	0.7724	0.6759	0.5908
Median	0.2897	0.2517	0.2253

TABLE III: Clinical Trials Results. All four submitted runs are compared against the best and averaged median performance on three measures: precision at 5, 10, and 15 documents.

Among our models, the pseudo-relevance feedback on 1000 results (MedIER_ct2) and the expanded set with 3000 results (MedIER_ct4) performs the best on average. However, the overall precision values are low. Our initial analysis showed that there were considerable number of topics in which no relevant documents were extracted by our runs. Additional error analysis is needed to check the root cause of this anomaly.

V. CONCLUSION AND FUTURE WORK

Our participation in this year's Precision Medicine track focused on the genetic variations mentioned in the query and explored ways to modify the search query and the retrieval ranking using this information. We also explored filtering retrieved clinical trial results based on demographic information. To achieve these, we built two indices for the scientific abstracts corpus – one on the free-text and the other on the MeSH identifiers mentioned in the metadata of the research articles. For the clinical trials corpus, we created one free-text index and an SQL database to store demographic eligibility criteria on age and gender extracted from the clinical trial descriptions. Corresponding to these indices, we created queries focusing on cancer types and gene names, augmented with a list of carefully selected synonyms; and a set of alternate queries on MeSH identifiers. The retrieval results from these queries were merged to get the final ranking. The results show that the merged results did better in pulling more relevant documents to the top of the ranked list (higher precision at 10), and that pseudo-relevance feedback improves the results even further. However, it is possible that using more flexible merging algorithms may boost the performance further. With additional training data, we would experiment with learning to rank algorithms and tuning parameters to boost the overall performance.

In the future, we plan to continue exploring improvements based on gene variant names for cancer-related document retrieval. Additional error analysis is needed to understand the key limitations in the experiments over the clinical trials corpus. In particular, we would analyze the accuracy of the demographic feature extraction component and its impact on the exclusion of valid clinical trials from our submitted runs.

REFERENCES

[1] Matthew S. Simpson, Ellen Voorhees, and William Hersh. Overview of the TREC 2014 Clinical Decision Support track. In *Proceedings of the 23rd Text Retrieval Conference (TREC)*, 2014.

[2] Kirk Roberts, Matthew S. Simpson, Ellen Voorhees, and William R. Hersh. Overview of the TREC 2015 Clinical Decision Support track. In *Proceedings of the 24th Text Retrieval Conference (TREC)*, 2015.

[3] Kirk Roberts, Dina Demner-Fushman, Ellen Voorhees, and William R. Hersh. Overview of the TREC 2016 Clinical Decision Support track. In *Proceedings of the 25th Text Retrieval Conference (TREC)*, 2016.

[4] Harsha Gurulingappa, Luca Toldo, Claudia Schepers, Alexander Bauer, and Gerard Megaro. Semi-supervised information retrieval system for clinical decision support. In *Proceedings of the 25th Text Retrieval Conference (TREC)*, 2016.

[5] Yanshan Wang, Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, Sijia Liu, and Hongfang Liu. An ensemble model of clinical information extraction and information retrieval for clinical decision support. In *Proceedings of the 25th Text Retrieval Conference (TREC)*, 2016.

[6] Simon Greuter, Philip Junker, Lorenz Kuhn, Felix Mance, Virgile Mermet, Angela Rellstab, and Carsten Eickhoff. ETH Zurich at TREC Clinical Decision Support 2016. In *Proceedings of the 25th Text Retrieval Conference (TREC)*, 2016.

[7] Brendan Kish, Thomas Walsh, Katherine Small, Steven Gassert, Kylie Small, and Sharon Gower Small. Siena's clinical decision assistant with machine learning. In *Proceedings of the 25th Text Retrieval Conference (TREC)*, 2016.

[8] Seung-Hyeon Jo and Kyung-Soon Lee. CBNU at TREC 2016 Clinical Decision Support Track. In *Proceedings of the 25th Text Retrieval Conference (TREC)*, 2016.

[9] Sadid A. Hasan, Siyuan Zhao, Vivek V. Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Aaditya Prakash, and Oladimeji Farri. Clinical question answering using key-value memory networks and knowledge graph. In *Proceedings of the 25th Text Retrieval Conference (TREC)*, 2016.

[10] Sarvnaz Karimi, Sara Falamaki, and Vincent Nguyen. CSIRO at TREC Clinical Decision Support Track. In *Proceedings of the 25th Text Retrieval Conference (TREC)*, 2016.

[11] Danchen Zhang, Daqing He, Sanqiang Zhao, and Lei Li. Query expansion with automatically predicted diagnosis: iRiS at TREC CDS Track 2016. In *Proceedings of the 25th Text Retrieval Conference (TREC)*, 2016.

[12] Fengmin Hu, Danny T.Y. Wu, Qiaozhu Mei, and V.G.Vinod Vydiswaran. Learning from medical summaries: The University of Michigan at TREC 2015 Clinical Decision Support Track. In *Proceedings of the 24th Text REtrieval Conference (TREC)*, 2015.