MayoNLPTeam at the TREC 2018 Precision Medicine Track: Simple Information Retrieval Approach Is the Best

Yanshan Wang
Mayo Clinic
Rochester, MN
wang.yanshan@mayo.edu

Andrew Wen
Mayo Clinic
Rochester, MN
wen.andrew@mayo.edu

Sijia Liu
Mayo Clinic
Rochester, MN
liu.sijia@mayo.edu

Hongfang Liu
Mayo Clinic
Rochester, MN
liu.hongfang@mayo.edu

ABSTRACT

This paper describes the participation of the Mayo Clinic NLP team in the Text REtrieval Conference (TREC) 2018 Precision Medicine (PM) track. The TREC 2018 PM track repeats the TREC 2017 PM track on retrieving relevant biomedical articles or clinical trials to cancer-related topics. We tested three information retrieval (IR) approaches in our official submission, including a simple approach of matching keywords and MeSH terms, an approach of matching extracted and normalized medical concepts, and a Learning-To-Rank approach based on TREC 2017 PM training data. In our systems, we used the NCI thesaurus and COSMIC to expand disease and gene terms with synonyms, respectively. Submissions were evaluated by the standard TREC test collections. Evaluation results show that our submissions using the simple IR approach have the best performance for both biomedical article and clinical trial retrieval subtasks. Recalling that our submissions using pseudo relevance feedback and Markov Random Field information retrieval models are also inferior to those using simple IR approaches in the TREC 2017 PM track, we conclude that the IR approaches shown effective in the general domain are not generalizable whilst retaining good performance for this medical track and the simple IR approach using keyword matching has the best record for consistent performance.

KEYWORDS

information retrieval; natural language processing; learning to rank; information extraction; precision medicine

1 INTRODUCTION

The Text REtrieval Conference (TREC) 2018 Precision Medicine (PM) track [5] repeats the TREC 2017 PM track [6] on retrieving useful precision medicine-related information to clinicians treating cancer patients. Given cancer patient topics, participants are challenged with two subtasks in this track: to retrieve biomedical articles in the form of article abstracts (largely from MEDLINE/PubMed), and to retrieve the clinical trials (from ClinicalTrials.gov) for which the patient is eligible. Similarly to the TREC 2017 PM track, topics are synthetic cases created by precision oncologists at the University of Texas MD Anderson Cancer Center. Each topic includes information on the patient's disease (type of cancer), the relevant genetic variants (which genes), and basic demographic information (age, sex). Table 1 shows a topic example. More details about the PM tracks can be found in the overview papers [5, 6].

Last year, our systems submitted to the TREC 2017 PM track tested multiple IR approaches, including using pseudo relevance feedback to expand query terms, using natural language processing (NLP) extracted entities (e.g., gene, variant, and disease) to re-rank retrieved documents, and using Markov Random Field as retrieval models [7]. The evaluation results, however, showed that these approaches failed to improve the retrieval performance and underperformed when compared with the simple baseline approach. Top performing systems utilized a variety of knowledge bases to carefully expand gene, variant, and disease terms, and had hand-crafted rules to pre- and post-process retrieved documents [2, 4]. Table 2 summarizes a number of knowledge bases used by the participating systems in the TREC 2017 PM track. Moreover, a set of hand-crafted rules were also applied in these top systems to pre- and post-process the documents to be indexed and documents being retrieved.

Table 1: An example of cancer patient topics in the TREC 2018 PM track. "melanoma" is the cancer name, "BRAF" is the gene name, "V600E" is the variant name, "64" is the age, "male" is the sex.

```
<topic number="1">
<disease>melanoma</disease>
<gene>BRAF (V600E)</gene>
<demographic>64-year-old male</demographic>
</topic>
```

Table 2: Knowledge bases used by the participating systems in the TREC 2017 PM track.

Knowledge base	How it was used
NCBI GeneDB	gene name expansion; find relevant PubMed articles
UMLS	ontological expansion
HGNC	gene name expansion
COSMIC	gene expansion; variant expansion
DrugBank	pre-annotation
NCI thesaurus	disease expansion
Entrez Gene Library	gene name expansion
MEDLINE	disease name expansion
MeSH hierarchy	disease name expansion
MeSH ontology codes	filter non-cancer articles
SNOMED/Lexigram	disease expansion
NCBI Homo Sapiens	gene expansion
FDA labels	build knowledge graphs
DGIdb	build knowledge graphs
SNOMED CT	use hypernyms for disease name expansion
PMDG	gene name expansion

Although the top systems have promising performance, they tend to use the conventional IR approaches and tweak the system using rules and knowledge bases that are usually specific to the systems rather than propose novel IR approaches that are reproducible and generalizable. The goal of our study is to test novel or existing IR approaches and to make our result and conclusion generalizable to other medical and clinical IR tasks. Therefore, in the TREC 2018 PM track we tested three IR approaches in our official submissions to both biomedical article and clinical trial retrieval subtasks, including a simple approach of matching keywords and MeSH terms, a approach of matching extracted and normalized medical concepts, and a Learning-To-Rank approach using TREC 2017 PM topics and gold standard as training data.

2 METHODS

2.1 Indexing

We utilized the open-source package Elasticsearch[1] as the platform for our IR system development.

For the biomedical article retrieval subtask, we indexed all the fields in the xml files without preprocessing. In addition, we indexed disease entities extracted by our Cohort Retrieval Enhanced by Analysis of Text (CREATE) pipeline [3], which maps conditions into Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) normalized forms. The CREATE pipeline is publicly available on GitHub[2]. We also indexed gene and disease entities annotated by PubTator [8], which is a Web-based tool developed by the NCBI to automatically annotate biological entities through the use of advanced text-mining techniques. The PubTator API is also publicly available[3].

For the clinical trial retrieval subtask, we indexed the clinical trials that were related to the cancers in the given topics. The detailed steps are shown in Table 3. In our systems, we mainly used two knowledge bases, the NCI thesaurus and COSMIC, for expanding disease and gene terms with synonyms, respectively. Some fields that contain only values, such as "minimum_age" and "maximum_age", were indexed as integer data type for the convenience of age comparison in the retrieval.

2.2 Simple Approach

For both subtasks, we submitted separate runs using a simple IR approach based on keyword and MeSH term matching.

For the biomedical article retrieval task, we queried the indexed "title" and "abstract" fields by matching keywords of disease (e.g., "lung cancer"), gene (e.g., "braf"), variant (e.g., "v600e"), descriptive gene information (e.g., "loss of function"), and gender (e.g., "male") in the topics, and queried the "MeSH Heading" field by matching the mapped MeSH terms of disease and gene terms in the topics. We utilized the NLM Medical Text Indexer (MTI) [1] to map the disease and gene terms into MeSH terms. MTI API is publicly available[4]. We also transformed age into MeSH terms based on the age range (e.g., "1-year-old" to "infant"), and queried the "MeSH Heading" field.

[1]https://www.elastic.co/products/elasticsearch
[2]https://github.com/OHNLPIR/OMOP_CDM_IO
[3]https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/tutorial/index.html
[4]https://ii.nlm.nih.gov/MTI/

Table 3: Steps of indexing clinical trials.

Step	Rule
1	Extract disease, gene, and variant terms from topics;
2	Use the NCI thesaurus to find synonyms to diseases, and COSMIC to find synonyms to genes and variants;
3	Filter out non-cancer clinical trials using the NCI thesaurus;
4	If text in the condition field of clinical trials contain the disease or disease synonyms in the topic, go to steps 5-6;
5	Index the matched topic disease in an extra field <topic_disease_mapping> and the matched topic disease synonyms in an extra field <topic_NCI_syn>;
6	If the topic gene or gene synonyms are matched in clinical trials but not negated and not in the exclusion criteria field, index the matched topic gene in an extra field <topic_gene_mapping> and the matched topic gene synonyms in an extra field <topic_cosmic_syn>;
7	If the maximum age in clinical trials is 'N/A', index 150; and if in the minimum age in clinical trials is 'N/A', index 0.

For the clinical trial retrieval task, we applied similar methodology and queried the "official_title", "brief_summary", and "mesh_term" fields. Unlike biomedical articles, clinical trials have inclusion and exclusion criteria of age and gender for patient recruitment. We therefore retrieved the documents with "gender" either matching the gender in the topic or equal to "all', and ensured that the age in the topic was within the range specified by the "minimum_age" and "maximum_age" fields. This is easy to implement using range query in Elasticsearch.

2.3 Approach of Matching Extracted and Normalized Medical Concepts

In this approach, we added matching of normalized disease terms to the simple approach. These normalized terms were extracted using the CREATE pipeline [3].

2.4 Learning-To-Rank

Topics are composed of disease, gene, variant, age, and sex, which are used to query different indexed fields in the simple IR approach. The Learning-To-Rank approach attempts to determine and assign higher weights to the information that are considered to be more important to physicians. Using the TREC 2017 PM topics and gold standard, weights could be computed by maximizing the retrieval metric, mean average precision (MAP). Figure 1 depicts the searching mechanism and weight parameters for the biomedical article retrieval. Unlike the simple IR approach, we also used synonyms for diseases and genes from the NCI thesaurus[5] and COSMIC[6], respectively. The loss function was defined as $1 - \text{MAP}$. Line search

[5]https://evs.nci.nih.gov/ftp1/NCI_Thesaurus/
[6]https://cancer.sanger.ac.uk/cosmic/download

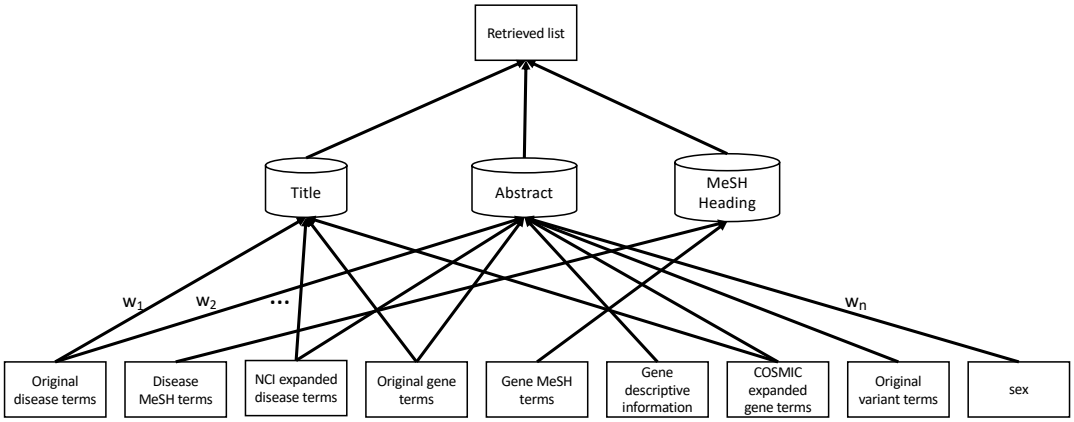


Figure 1: Learning-To-Rank for the biomedical article retrieval.

Table 4: A method summary of official submissions.

Task	Run name	Keyword matching	MeSH term matching	CDM concept matching	Learning-To-Rank	Pubtator
Biomedical article retrieval	medsimp	•	•			
	medcreat	•	•	•		
	medcomp	•	•		•	
	pubtator	•	•			•
Clinical trial retrieval	mayoctsimp	•	•			
	mayoctscreat	•	•	•		
	mayoctcomp	•	•		•	

was utilized as the training algorithm. The same Learning-To-Rank approach was also applied for the clinical trial retrieval subtask.

2.5 Pubtator

In a separate submission specifically for the biomedical article retrieval task, we utilized the extracted gene and disease entities annotated by PubtTator [8]. PubTator is a Web-based tool that pre-annotates biomedical entities (e.g., gene, disease, mutation) from the entire content of PubMed. It combines a set of tools that have been extensively evaluated to ensure high quality of automatically extracted results. Since Pubtator normalized gene and disease terms in the PubMed articles into the NCBI Gene and MEDIC format, we also normalized gene and disease terms in the topics into the same format so as to be able to match them in the index. We additionally utilized the age and sex information in the same manner as the simple IR approach.

2.6 Official submissions

Table 4 summarizes the official submitted runs to the PM track.

3 RESULTS

Figure 2 plots the training curve of the Learning-To-Rank approach based on the TREC 2017 PM training dataset. It shows that the MAP increases along iterations and becomes maximal with the optimal weight parameters. Figures 3 and 4 depict heat maps of the optimal weight parameters for the biomedical article and clinical trial retrieval, respectively. First, we can observe that the model results in genes having the highest weight among the available information in the topics, indicating that gene matches have the highest impact on relevance. For the biomedical article retrieval subtask, searching original disease and gene terms in the title field, and searching original gene terms and gene descriptive information in the abstract were the most weighted for retrieving relevant documents. Searching original variant terms and sex in the abstract field was the second most weighted. The MeSH terms along with expanded disease and gene terms using NCI thesaurus and COSMIC are just slightly more weighted than the remainder of the available information. For the clinical trial retrieval subtask, we did not consider expanding disease and gene terms in the topics as clinical trial documents were pre-processed prior to indexing and disease and gene terms were normalized to match those that would appear in the topics. We can obviously see from the heat map that the gene and sex information is the most important for this subtask. Results from both subtasks show that gene information is crucial to the PM retrieval task.

Submissions were evaluated by the standard TREC test collections. Table 5 lists the overall results of our official submissions. Surprisingly, the simple approach (Runs *medsimp* and *mayoctsimp*) achieved the best performance for both subtasks in terms of infNDCG. The Learning-To-Rank approach (Runs *medcomp* and

Table 5: Overall results of official submissions.

Task	Run name	infNDCG	P@10	R-prec
Biomedical article retrieval	medsimp	**0.4036**	**0.56**	**0.2726**
	medcreat	0.2229	0.302	0.1352
	medcomp	0.3891	0.51	0.2568
	pubtator	0.3695	0.488	0.2108
Clinical trial retrieval	mayoctsimp	**0.4324**	0.446	0.3219
	mayoctscreat	0.4183	0.456	0.3213
	mayoctcomp	0.4281	**0.506**	**0.3235**

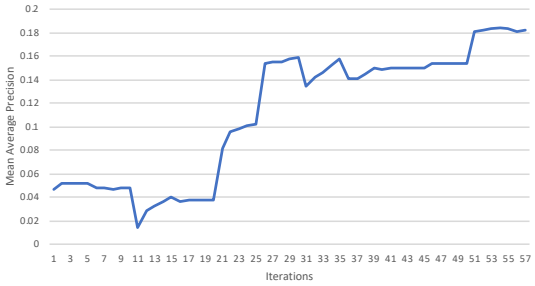


Figure 2: The training curve of MAP in the Learning-To-Rank based on the TREC 2017 PM training dataset.

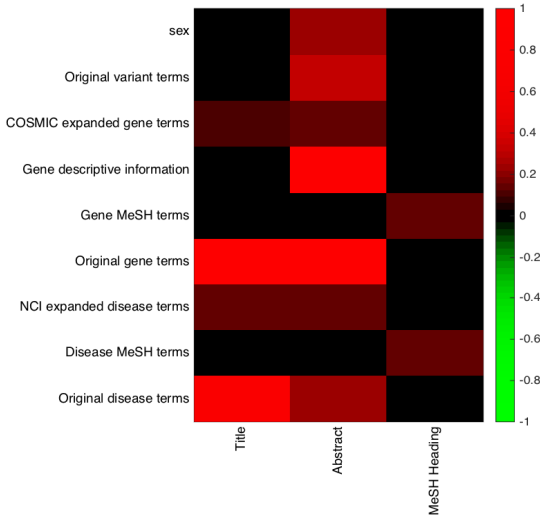


Figure 3: A heat map of weight parameters for the biomedical article retrieval.

mayoctcomp) is only slightly better for the clinical trial retrieval subtask in terms of P@10 and R-prec. This result indicates that the learning algorithm ranks relevant documents higher in the retrieved list and leads to better P@10's. However, our results show that this conclusion is not true in the biomedical article retrieval. The approach of matching extracted and normalized medical concepts (Runs *medcreat* and *mayoctcreat*) have the worst performance,

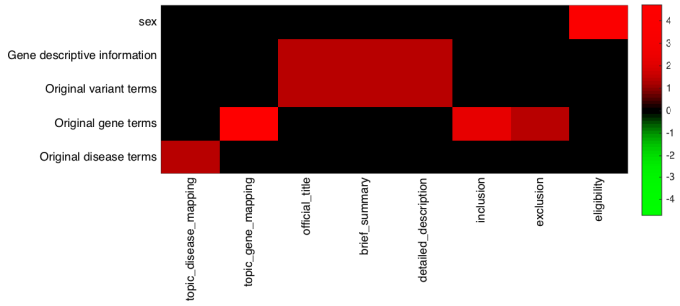


Figure 4: A heat map of weight parameters for the clinical trial retrieval..

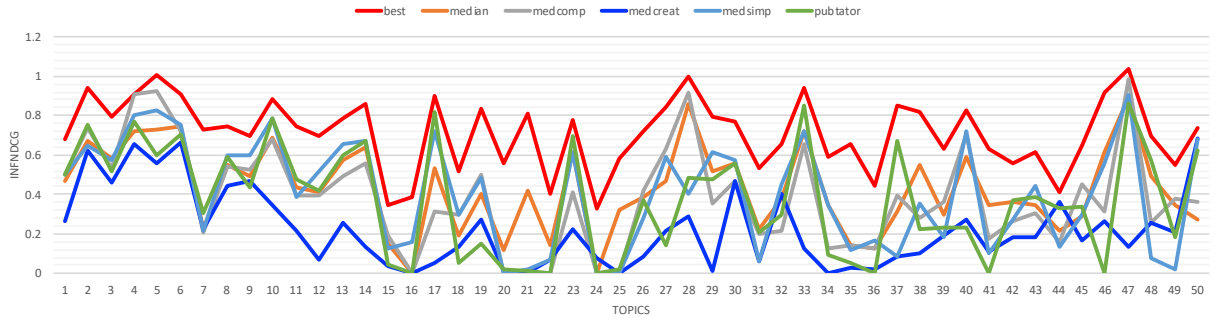
which may be due to the high reliance on the performance of concept extraction and normalization. The results of PubTator (Run *pubtator*) are competitive.

Figures 5 and 6 compare our submissions with the best and median results for each topic. We can see that our systems perform better if measured by P@10 rather than by infNDCG for both subtasks. For example, the Learning-To-Rank approach has the best P@10 among all submitted systems for 5 out of 50 topics for the biomedical article retrieval and 10 out of 50 topics for the clinical trial retrieval. It has the best infNDCG only for 1 topic for the first subtask and 3 topics for the second subtask. The simple approach also has promising P@10's and it outperforms all submitted systems for 8 topics for the first subtask and 7 topics for the second subtask. Overall, the simple approach is superior to the Learning-To-Rank approach.

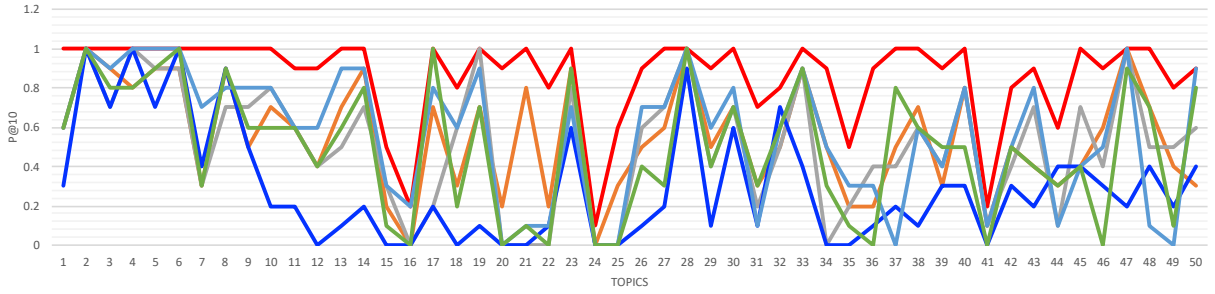
4 CONCLUSION AND DISCUSSION

This paper describes our participation in the biomedical article retrieval and clinical trial retrieval subtasks of the TREC 2018 PM track. Our official submissions are based on three IR approaches, including a simple IR approach of matching keywords and MeSH terms, an approach of matching extracted and normalized medical concepts, and a Learning-To-Rank approach based on TREC 2017 PM training data. The evaluation results show that our submissions using the simple IR approach have the best performance for both subtasks.

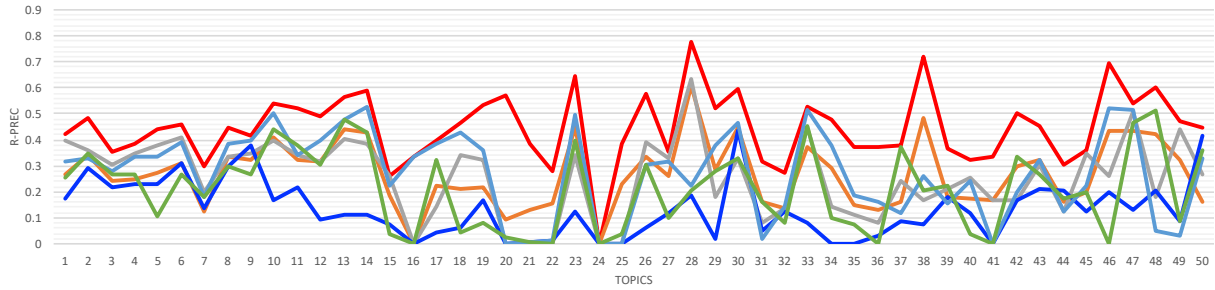
Recalling that our submissions using pseudo relevance feedback and Markov Random Field information retrieval models are also inferior to those using simple IR approaches in the TREC 2017



(a) infNDCG



(b) P@10



(c) R-prec

Figure 5: Results per topic of the biomedical article retrieval.

PM track, we conclude that the IR approaches shown effective in the general domain are not generalizable whilst retaining good performance for this medical track and the simple IR approach using keyword matching has the best record for consistent performance.

Most top performing systems of the TREC 2017 PM track incorporate many knowledge bases and resources into the IR systems and utilize heuristics rules to pre- and post-process topics and documents. However, these rules and the usage of knowledge bases are usually specific to those top performing systems, which lack reproducibility and generalizability. Since those systems heavily rely on the knowledge bases and heuristic rules, the PM track tends to encourage IR researchers to tweak the system rather than study the fundamental IR models that are reproducible and generalizable to other medical and clinical IR tasks. Therefore, we encourage the IR community to study reproducible approaches of using knowledge bases and rules and to study fundamental IR models that are generalizable and reusable to the IR tasks in the medical and clinical domain.

ACKNOWLEDGMENTS

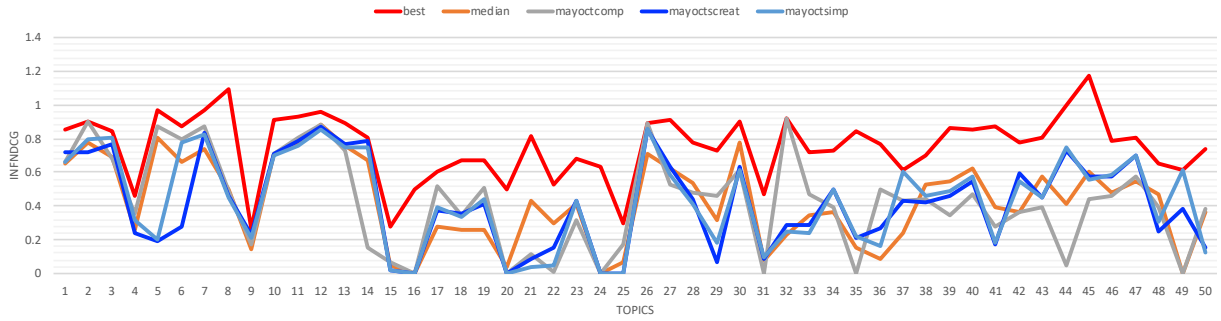
This work has been supported by the National Institute of Health (NIH) grants R01LM011934, R01GM102282, R01EB19403, R01LM11829 and U01TR002062.

REFERENCES

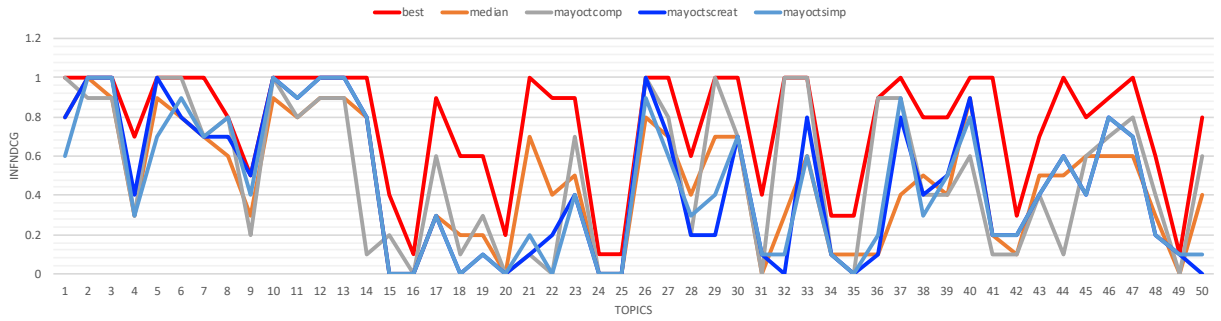
[1] Alan R Aronson, James G Mork, Clifford W Gay, Susanne M Humphrey, Willie J Rogers, et al. 2004. The NLM indexing initiative's medical text indexer. *Medinfo* 89 (2004).

[2] Travis Goodwin, Michael Skinner, and Sanda Harabagiu. 2017. UTD HLTRI at TREC 2017: Precision Medicine Track. In *TREC*. Gaithersburg, MD.

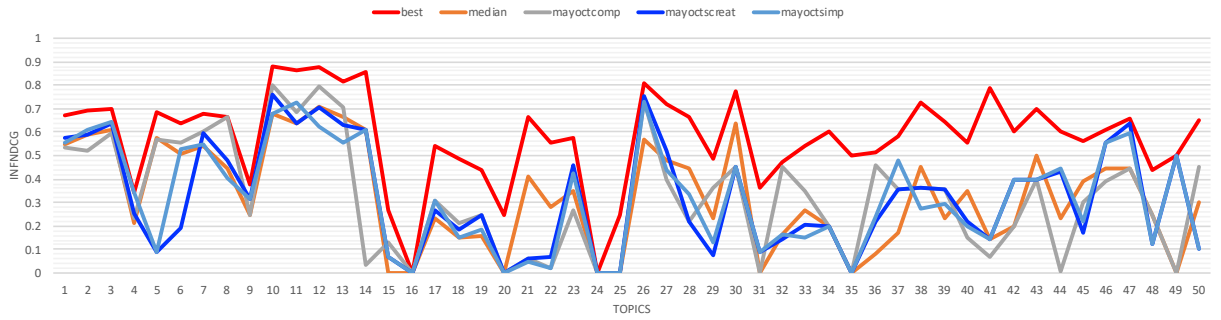
[3] Sijia Liu, Yanshan Wang, Andrew Wen, Liwei Wang, Na Hong, Feichen Shen, Steven Bedrick, William Hersh, and Hongfang Liu. 2018. CREATE: Cohort Retrieval Enhanced by Analysis of Text from Electronic Health Records using OMOP Common Data Model. (2018).



(a) infNDCG



(b) P@10



(c) R-prec

Figure 6: Results per topic of the clinical trial retrieval.

[4] ASM Ashique Mahmood, Gang Li, Shruti Rao, Peter McGarvey, Cathy Wu, Subha Madhavan, and K Vijay-Shanker. 2017. UD_GU_BioTM at TREC 2017: Precision Medicine Track. In *TREC*. Gaithersburg, MD.

[5] Kirk Roberts. 2018. Overview of the TREC 2018 precision medicine track. In *TREC*. Gaithersburg, MD.

[6] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. 2017. Overview of the TREC 2017 precision medicine track. In *TREC*. Gaithersburg, MD.

[7] Yanshan Wang, Ravikumar Komandur-Elayavilli, Majid Rastegar-Mojarad, and Hongfang Liu. 2017. Leveraging both Structured and Unstructured Data for Precision Information Retrieval. In *TREC*. Gaithersburg, MD.

[8] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research* 41, W1 (2013), W518–W522.