CSIRO at 2019 TREC Precision Medicine Track

Maciej Rybinski
CSIRO Data61
Marsfield, NSW, Australia
maciek.rybinski@csiro.au

Sarvnaz Karimi
CSIRO Data61
Marsfield, NSW, Australia
sarvnaz.karimi@csiro.au

Cecile Paris
CSIRO Data61
Marsfield, NSW, Australia
cecile.paris@csiro.au

ABSTRACT

TREC Precision Medicine track focuses on search tasks tailored for oncologists. Given a cancer patient, the proposed systems must find clinical trials that match the patient, as well as the relevant information from biomedical literature (PubMed abstracts 2019 baseline). In our experiments, we compare BM25 and Divergence from Randomness (DFR) baselines and report results obtained with multiple learning-to-rank models. Some of our submitted runs score in top ten runs reported by the organisers.

1. INTRODUCTION

Precision Medicine is the development of treatment plans that take into account the patients' unique genetic markup, environmental influences, and lifestyle choices, as well as other biomarker information for an individual's prevention, diagnosis, and treatment strategies [4].

The TREC Precision Medicine (PM) track [8, 9], a specialisation of the TREC Clinical Decision Support track [10, 7, 11], aims to tackle the challenge of including cancer genetic information in designing treatment strategies. It aims to provide the medical staff clinical decision support for cancer patients. The task is its third year and requires participants to develop search systems that retrieve relevant biomedical literature and clinical trials for decision support, given a query with the patient's genetic mutations, disease, and demographic attributes.

In this report, we outline our CSIROmed team submission for the 2019 PM track, discuss the experimental setup and present our results. We also report on runs that corrected our submission errors, which led to improvements over our submitted runs. Overall, we show benefit in applying learning-to-rank using some of the conventional methods that do not rely on large amount of training data. In none of our runs any hand-crafted rules is applied.

2. DOCUMENTS AND TOPICS

In TREC Precision Medicine 2019, two datasets are used: (1) biomedical literature of approximately 29.1 million PubMed journal abstracts (a December 2018 MEDLINE snapshot); and, (2) May 2019 snapshot of ClinicalTrials.gov comprising of 306, 238 clinical trials.

Forty topics in a semi-structured format are provided. Each topic contains disease (a cancer type, gene(s) or genetic mutation(s) specific to the patient and related to their condition, and demographic attributes of the patient. That is, each topic is representative of one patient, as created by precision oncologists at the University of Texas MD Anderson Cancer Center. For example, the case of a 58 y/o male with gastrointestinal stromal tumor with KIT gene exon 9502_503 duplication is shown in Figure 1.

```
<topic number="9">
<disease>
gastrointestinal stromal tumor
</disease>
<gene>
KIT (exon 9 502_503 duplication)
</gene>
<demographic>58−year−old male</demographic>
</topic>
```

Figure 1: A TREC 2019 precision medicine topic (Topic 9).

3. RETRIEVAL SYSTEM

We use the same document parsing and indexing approach that we developed previously as part of our participation in Clinical Decision Support track 2016 [2, 3] (A2A system), Precision Medicine 2017 and 2018 [5, 6]. For indexing and processing of the documents (stemming, stopword elimination), we use the standard mechanisms of Apache Solr 6.6.2. This year we simplify query formulation process in all runs, using a concatenation of *gene* and *disease* fields.

In both sub-tasks, scientific abstracts and clinical trials, we report two baselines obtained with distinct retrieval models, BM25 and a divergence from randomness model—specifically InL2 or Inverse Document Frequency model with Laplace after-effect and normalisation 2—which we refer to as DFR, in order to compare their performance directly. All other experimental runs are focused on experimenting with different versions of our learning-to-rank approach. In case of clinical trials sub-task we also apply strict matching of patient's demographic attributes on all non-baseline runs.

Our re-ranking (learning-to-rank) approach is based on features designed to model query-document matches independently for *disease* and *gene* topic fields. For this purpose we use features based on similarity of word2vec centroids of respective query fields and phrases of the documents, which are re-scored. For the disease topic field the procedure is as follows. Contents of *disease* field are matched against phrases extracted from the document via a sliding window of varying size. Best match for each window size becomes a feature for our model. Goodness of a match is calculated as

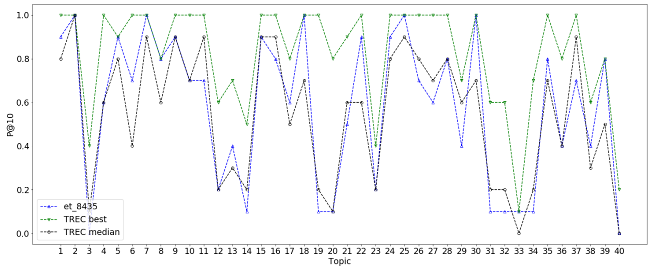

Figure 2: Per query comparison of our best run for abstracts versus the TREC best and median (P@10). Lines between the datapoints do not represent any results.

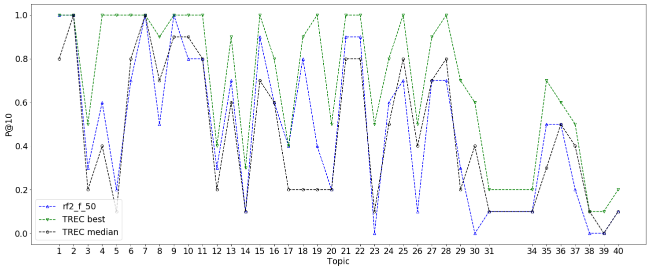

Figure 3: Per query comparison of our best run for clinical trials versus the TREC best and median (P@10). Lines between the datapoints do not represent any results.

cosine similarity between tf-idf weighted word-embeddings-based centroid representations of the field contents and extracted phrases. We use the sliding window sizes between 1–5; we use 30-dimensional skip-gram vectors trained on a 2018 MEDLINE snapshot and IDF data calculated on the same corpus. Apart from the 5 'best-match' features we use exact match for disease name (boolean), length of the disease topic field and length of the document.

Gene-related query-document matches are modeled in a similar fashion, so a top match is also extracted using IDF-weighted centroids and sliding windows of varying length. The feature extraction for gene information is designed for topic fields, which contain any number of gene mentions and some additional information (gene mentions and other information is parsed from the topic field contents). The features are calculated for specific gene names mentioned in the topic and topic parts in general (including the names). Parts of the gene field are fragments separated with a comma.

Gene-related features are: top matches for each window size (1–3) across all gene mentions, together with top match found across all window sizes (1–5), across all field parts. If the gene information is missing (for example, if the field is only 'high mutational burden'), we use the top-3 matches obtained for the field parts (in case of our example, the phrase 'high mutational burden').

Our models are trained directly on TREC PM 2017/2018 relevance judgements. For model selection we use randomized test and validation sets of three topics each and averaged the results over 100 of these randomized selections while re-training the models on the remaining topics. In the experiments, we use either *ExtraTrees* (a Random Forest variant) or XGBoost binary classifiers. In both cases, we treat their probabilistic outputs as relevance scores.

Our learning to rank approach follows a typical two-step setting. That is, we use our base ranker (see BM25 baselines) and re-rank a top portion of its results (50/100 top candidate documents). In our submitted runs we only reported the re-scored documents, which eventually resulted in poor inferred NDCG scores of these runs.

4. SUBMITTED RUNS

We submitted nine automatic runs for the two tasks, five runs on clinical trials and four runs on medical literature. Details of these runs are described in Table 1. We used equal weights for *disease* and *gene* terms.

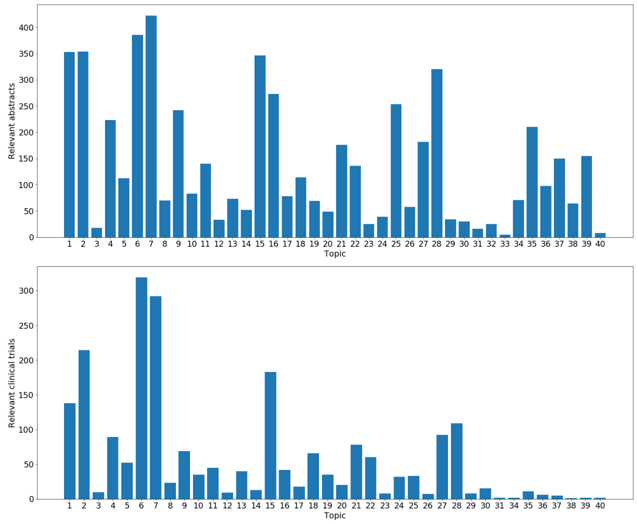

Figure 4: Number of relevant and partially relevant documents per query as listed in relevance judgements. Top bars are for abstracts and bottom is clinical trials.

5. RESULTS

An overview of our results is shown in Tables 2 and 3. The top row shows TREC Median results on 63 runs submitted by different teams for the scientific literature sub-task as reported in the TREC Overview [9]. A corresponding TREC Median results for clinical abstracts were calculated over 53 runs submitted by the participants. We also list the results of the team that had highest scores compared to other teams, submitted by JULIE Lab [1]. Their best run for the abstract retrieval is called `jlpmcommon2` and uses a learning-to-rank approach designed based on LETOR.

Clinical trial runs were different. There was no one run that achieved highest score for all three metrics. However, JULIE lab still scored highest in two of its runs for infNDCG and R-Prec, which indicates their underlying method is effective, with differences in the query processing methods for abstracts and clinical trials.

Learning-to-Rank. Our learning-to-rank runs based on ExtraTrees model seem to show encouraging results in terms of P@10, resulting in above-median scores both for clinical trials and scientific abstracts. Figures 2 and 3 present P@10

	Method			
Run	Ranking	Re-ranking	Re-ranking model	Demographic Filtering
Abstracts				
bm25_6801	BM25	–	–	–
dfr_9464	DFR	–	–	–
et_8435	BM25	Top 50	ExtraTrees	–
xgb_5113	BM25	Top 50	XGBoost	–
Clinical Trials				
bm25_ct_25	BM25	–	–	–
DFRInL2_f	DFR	–	–	✓
bm25_ct_f_61	BM25	–	–	✓
rf1_f_100	BM25	Top 100	ExtraTrees	✓
rf2_f_50	BM25	Top 50	ExtraTrees	✓

Table 1: Configuration of the CSIROmed submitted runs.

Run	infNDCG	P@10	R-Prec
TREC Median	0.4559	0.5450	0.2806
TREC Top Run	0.5783	0.6525	0.3572
Submitted			
bm25_6801	0.4553	0.5250	0.3029
dfr_9464	0.4531	0.5100	0.2948
et_8435	0.3239	0.5725	0.1856
xgb_5113	0.3111	0.5100	0.1796
Post-TREC			
bm25_6801	0.4693	0.5300	0.3093
dfr_9464	**0.4766**	0.5350	**0.3165**
et_8435	0.4727	**0.5825**	0.3092
xgb_5113	0.4592	0.5150	0.3032

Table 2: Search over abstracts. TREC Median is averaged over 40 topics.

Run	infNDCG	P@10	R-Prec
TREC Median	0.5137	0.4658	0.3477
TREC Top Run	0.6451	0.5947	0.4820
Submitted			
bm25_ct_25	0.4818	0.4632	0.3384
DFRInL2_f	0.4930	0.4684	0.3406
bm25_ct_f_61	0.4906	0.4658	0.3586
rf1_f_100	0.4450	0.4868	0.2880
rf2_f_50	0.3871	0.4921	0.2663
Post-TREC			
bm25_ct_25	0.5620	0.5000	0.4190
DFRInL2_f	**0.5842**	0.5132	**0.4320**
bm25_ct_f_61	0.5787	0.5053	0.4305
rf1_f_100	0.5064	0.4868	0.3529
rf2_f_50	0.5226	**0.5158**	0.3496

Table 3: Search over clinical trials. TREC Median is averaged over 40 topics.

per-query results in comparison with TREC median and best results, for scientific abstracts and clinical trials respectively. On this particular set of features the Random Forest variant seems to be a better fit than XGBoost. In both subtasks the baseline runs resulted in scores comparable to the median, with slightly below-median P@10 for scientific abstracts.

Learning-to-rank does not work well for topics with low precision at *k*, where we re-rank top-*k* documents–low numbers of relevant documents translates directly into poor performance. Number of relevant documents per-topic is presented in Figure 4 to provide context.

Post-TREC Runs. After submissions, we discovered two bugs in our our submitted runs: (1) with re-ranking, we only included the re-ranked portion of the results in the results files, which negatively impacted the scores; and (2) top documents from ranked results were chopped due to a bug in the code that did not include rank zero.

We include corrected runs in the second half of Tables 2 and 3 as *post-TREC*. For both abstracts and clinical trials, these runs lead to our best results in all three metrics. While they do not beat the top runs reported in the overview report, they would set our systems around third best on average for all three scores.

6. SUMMARY

Learning-to-rank models were the main focus of our experiments. We submitted runs for both *abstracts* and *clinical trials* document sets. Our best approaches for both document sets use ExtraTrees model for re-ranking top-50 documents, resulting in above-median infNDCG, P@10, and R-Prec scores.

Acknowledgements. Authors would like to acknowledge the funding provided by CSIRO Precision Health – Future Science Platform (FSP) to support our work. We also acknowledge the contributions of our colleagues over the years to develop the A2A underlying system.

References

[1] E. Faessler and M. Oleynik. JULIE lab at the 2019 TREC precision medicine track. In *TREC*, Gaithersburg, MD, 2019.

[2] S. Karimi, S. Falamaki, and V. Nguyen. CSIRO at TREC

clinical decision support track. In *TREC*, Gaithersburg, MD, 2016.

[3] S. Karimi, V. Nguyen, F. Scholer, B. Jin, and S. Falamaki. A2a: Benchmark your clinical decision support search. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, page 1277–1280, Ann Arbor, MI, USA, 2018.

[4] I. Konig, O. Fuchs, G. Hansen, E. von Mutius, and M. Kopp. What is precision medicine? *European Respiratory Journal*, 50(4), 2017.

[5] V. Nguyen, S. Karimi, S. Falamaki, D. M. Aliod, C. Paris, and S. Wan. CSIRO at 2017 TREC precision medicine track. In *TREC*, Gaithersburg, MD, 2017.

[6] V. Nguyen, S. Karimi, and B. Jin. An experimentation platform for precision medicine. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1357–1360, Paris, France, 2019.

[7] K. Roberts, D. Demner-Fushman, E. Voorhees, and W. Hersh. Overview of the TREC 2016 clinical decision support track. In *TREC*, Gaithersburg, MD, 2016.

[8] K. Roberts, D. Demner-Fushman, E. Voorhees, W. Hersh, S. Bedrick, and A. Lazar. Overview of the TREC 2018 precision medicine track. In *TREC*, Gaithersburg, MD, 2018.

[9] K. Roberts, D. Demner-Fushman, E. Voorhees, W. Hersh, S. Bedrick, A. Lazar, S. Pant, and F. Meric-Bernstam. Overview of the TREC 2019 precision medicine track. In *TREC*, Gaithersburg, MD, 2019.

[10] K. Roberts, M. Simpson, E. Voorhees, and W. Hersh. Overview of the TREC 2015 clinical decision support track. In *TREC*, Gaithersburg, MD, 2015.

[11] K. Roberts, M. S. Simpson, E. Voorhees, and W. R. Hersh. Overview of the trec 2015 clinical decision support track. In *Text REtrieval Conference*, Gaithersburg, MD, 2015.