CSIROmed at TREC Precision Medicine 2020

Maciej Rybinski Sarvnaz Karimi
CSIRO Data61
Sydney, Australia
firstname.lastname@csiro.au

ABSTRACT

TREC Precision Medicine (PM) focuses on providing high-quality evidence from the biomedical literature for clinicians treating cancer patients. Our experiments focus on incorporating *treatment* into search. We established a promising baseline using PM 2017-2018 datasets for training and 2019 for validation. Our baseline consisted of a base-ranking step using Divergence From Randomness (DFR) scoring that used disease and gene as queries and an aggregated text field to represent documents, followed by a BERT-based neural re-ranker. We examined two mechanisms for incorporating the treatment within the query formulation strategy for DFR: (1) a concatenation of disease, gene and treatment fields; and (2) a concatenation of disease and gene fields, but filtering out the documents where treatment terms were absent. We experimented with both strategies in combination with re-rankers trained either directly on TREC PM 2017-2019 retrieval task, or trained on a treatment-augmented version of these tasks.

We obtained the best results using boolean retrieval for treatment terms with a re-ranker trained on non-augmented TREC PM datasets. Our top-ranking run achieved 0.530, 0.565, 0.436 for infNDCG, P@10, RPrec, respectively. TREC median for these metrics were 0.432, 0.465, and 0.326.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; *Language models*; *Decision support systems*; • **Applied computing** → *Health informatics*.

KEYWORDS

Precision medicine; Search; Medical information retrieval; Learning-to-rank; Evidence-based medicine

ACM Reference Format:
Maciej Rybinski Sarvnaz Karimi. 2021. CSIROmed at TREC Precision Medicine 2020. In *TREC'20: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY.* ACM, New York, NY, USA, 3 pages.

TREC 2020, November, 2020, Online

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00

1 INTRODUCTION

TREC Precision Medicine (PM) focuses on providing high-quality evidence from the biomedical literature for clinicians treating cancer patients. In 2020, TREC PM aimed at search that provides evidence on pros and cons of a given treatment for a cancer patient from the published literature [4]. That is, it simulated a scenario where an oncologist seeks for information on a treatment, given a type of cancer and genetic mutation(s). We participated in the TREC PM as CSIROmed team, experimenting with neural ranking for precision medicine. We experimented with a two-step information retrieval framework, consisting of initial ranking step and a re-ranking step. For the initial ranking we use an established, strong, word-matching baseline—Divergence from Randomness (DFR) [1]—using information content modeled with inverse document frequency. The initial ranking step is followed by a re-scoring of top 100 results using a neural re-ranker, obtained through fine-tuning a BioBERT model on historical TREC PM datasets (2017-2018) [3, 5].

We analyse several approaches for incorporating the treatment aspect—new in TREC Precision Medicine 2020 shared task—in the search system pipeline. We evaluate two simple ways of incorporating the treatment aspect within the initial ranking step (adding the treatment term to a disjunctive query versus boolean filtering of the results based on presence of the treatment term) in combination with two different flavours of the re-ranker. We consider re-ranking using:

(1) direct fine tuning on historical TREC PM data; and,
(2) using a treatment-augmented dataset obtained from the original TREC PM data.

2 DATASETS

Search systems were evaluated with a collection of 31 topics, each with disease, gene and treatment fields e.g., (Disease: colorectal cancer, Gene: ABL1, Treatment: Regorafenib). Each topic represents a patient profile. The corpus used for search is a 2019 PubMed snapshot.

3 METHODS

Overview of the search pipeline. As already mentioned, we present the experiments within a 2-step search framework (ranking/re-ranking), consisting of: (1) an initial retrieval step using DFR, and (2) neural re-ranking with BioBERT for the top 100 documents of the initial ranking.

Queries are formulated from the topics independently at each of the two steps.

Indexing and initial retrieval step with DFR. We index the Medline 2019 snapshot using the following fields: *PMID* (PubMedID), *title*, *abstract*, *author-given keywords*, *MeSH descriptors*, *MeSH qualifiers*. An inverted index representation is also created for an aggregated *text* field.

For the initial retrieval step we use a DFR model with information content modeled with inverse document frequency with Laplacian after-effect and H2 normalisation.

We evaluate two query formulation strategies: (DIS) a simple disjoint query, formulated with disease, gene and treatment topic fields; and (FIL) a disjoint query over disease and gene terms with a Boolean filter for presence of the treatment term(s).

Neural re-ranking with BioBERT. We produce a the (neural) re-ranking relevance score using an output of a fine-tuned BioBERT with binary linear layer connected to the encoder's pooled layer with dropout. BioBERT is a domain-adapted BERT variant with additional pre-training on an earlier snapshot of the PM abstract corpus [2]. The re-ranking model is fine-tuned using cross-entropy loss and pointwise re-ranking approach, so we essentially train a binary classifier on binarised human judgments from 2017—2019 TREC PM datasets. For inference (for search) we use a softmax over the classifier output as BERT-based relevance score.

Training on human judgment means that a training instance is a topic (query) and document pair. We use a queries formulated as concatenations of specific topic fields, (e.g., *disease* and *gene*) as BERT's 'Sentence A' input and document representation (title and abstract) as 'Sentence B' input.

Treatment Augmentation. We experiment with 'treatment-augmentation' of the 2017—2018 TREC PM training data to increase the value of the historical data in fine-tuning the BioBERT re-ranker for the TREC 2020 task (so, covering the 'new' treatment aspect).

For this purpose we use a list of drug names from DrugBank and cross-reference it with MeSH terms and keywords of articles from the human judgment data from the past tracks.

To create treatment-augmented judgment dataset we iterate through the (PMID, topic number, relevance score) triples. For each triple we check, if any of the keyword/MeSH terms of the document corresponding to a given PMID appears in the DrugBank name list. For each keyword (i.e., treatment) that does, we add a 'treatment-augmented' quad to the augmented dataset. Each quad of this augmented dataset has PMID, topic number, relevance score and treatment (all but the treatment 'inherited' from the original triple).

After the initial augmented dataset is compiled, for each positive quad added, we add a negative quad with randomly sampled treatment expression. This is to ensure that the model can learn the difference between a hit and a miss on the treatment aspect.

We use the augmented dataset as if the treatment expressions of each of the quads pertained to the respective topics. The specifics of query formulation for re-ranking are presented below.

Query formulation for re-ranking. We evaluate two variants of fine-tuning, each of them used differently at re-ranking stage as well. The first variant is fine-tuned directly on the non-augmented human judgments of the 2017-2019 PM datasets. In this variant (hereafter referred to as 'non-augmented') the query (so, the 'Sentence A' input) is formulated as a space-separated concatenation of *disease* and *gene* topic fields, both for tuning and actual re-ranking.

In the variant using treatment-augmented data (referred to as 'augmented'). The 'Sentence A' inputs are formulated as: 'disease: D, gene: G, treatment: T', where D and G are *disease* and *gene* topic fields respectively. T denotes the treatment added to training quads in the augmentation process (in training), or the *treatment* topic field (for re-ranking).

4 EXPERIMENTS

In our runs submitted for evaluation in TREC we experiment with different combinations of DIS/FIL (initial retrieval) and augmented/non-augmented re-ranking. We submitted DIS-augmented, DIS-non-augmented, FIL-augmented and FIL-non-augmented runs, as well as a FIL-baseline (with no reranking). Here, we also include results of a DIS-baseline (no reranking).

FIL runs in our TREC submission had tail documents appended from corresponding DIS runs to avoid returning less than 1000 documents per topic. It means that for each topic we appended documents non-present in the original FIL-baseline ranking in the order they appeared in DIS-baseline, until the 1000 document limit was reached. The same procedure was performed for FIL-augmented/DIS-augmented and FIL-non-augmented/DIS-non-augmented respectively. Our post TREC experiments revealed that the merging procedure had minimal impact on the effectiveness of the FIL runs.

We report search effectiveness metrics as reported by TREC organisers: P@10, infNDCG and RPrec.

5 RESULTS AND DISCUSSION

We obtained the best results using boolean retrieval for treatment terms with a re-ranker trained on non-augmented TREC PM datasets (FIL-non-augmented). Our best run achieved 0.530, 0.565, 0.436 for infNDCG, P@10, RPrec, respectively. TREC median for these metrics were 0.432, 0.465, and 0.326. In the official evaluation of TREC Precision Medicine 2020 FIL-non-augmented was the best performing run in R-prec and P@10, and second best in infNDCG. The results for all runs are presented in Table 1.

Although the augmented runs have not led to improvements over the baseline, the comparison between DIS-augmented and DIS-non-augmented suggests that supervised training on augmented data results in a re-ranking model adapted to the new search aspect. FIL-augmented outperforms FIL-non-augmented, because the treatment aspect is incorporated restrictively in the initial retrieval step and FIL-augmented does better in covering the remaining two aspects (disease and gene).

Method	TREC run	**infNDCG**	**R-prec**	**P@10**
DIS-baseline	–	0.502	0.380	0.523
DIS-augmented	rRRa	0.493	0.373	0.532
DIS-non-augmented	rlxRR	0.475	0.378	0.484
FIL-baseline	strDFR	0.513	0.412	0.526
FIL-augmented	sRRa	0.527	0.412	0.529
FIL-non-augmented	strRR	**0.530**	**0.436**	**0.565**
TREC Median		0.432	0.465	0.326

Table 1: Results of our runs and comparison with TREC median (over 66 runs, 16 teams). DIS: disjoint query formulated with disease, gene and treatment topic fields. FIL: disjoint query over disease and gene terms with a Boolean filter for presence of the treatment term(s).

6 CONCLUSIONS

The main theme of our submission to the TREC PM 2020 track was to experiment with neural re-ranking focusing on task adaptation for this year's problem, which includes a new search aspect (treatment) as compared to previous years. We obtain the best results including the treatment aspect through Boolean filtering and following the initial retrieval step with a neural re-ranker trained directly on the past TREC PM relevance judgments. This led our team achieve the top ranking run among all teams for R-Prec and P@10.

We also propose a treatment-augmentation approach, in which we reformulate the training data to resemble the new task (using external resources). Although we achieve best results with a non-augmented model, the augmentation procedure shows enough promise in aspect-adaptation to make it a suitable avenue for future work.

REFERENCES

[1] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *TOIS*, 20(4):357–389, 2002.

[2] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[3] K. Roberts, D. Demner-Fushman, E. Voorhees, W. R. Hersh, S. Bedrick, A. Lazar, and S. Pant. Overview of the TREC 2017 Precision Medicine track. In *TREC*, Gaithersburg, MD, 2017.

[4] K. Roberts, D. Demner-Fushman, E. M. Voorhees, S. Bedrick, and W. R. Hersh. Overview of the TREC 2020 Precision Medicine Track. In *TREC*, 2020.

[5] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, and A. J. Lazar. Overview of the TREC 2018 Precision Medicine Track. In *TREC*, Gaithersburg, MD, 2018.