Filter, Transform, Expand, and Fuse
The IMS Unipd at TREC 2021 Clinical Trials

Giorgio Maria Di Nunzio, Guglielmo Faggioli, Stefano Marchesin

Department of Information Engineering
University of Padua, Italy
{giorgiomaria.dinunzio, stefano.marchesin}@unipd.it,
{guglielmo.faggioli}@phd.unipd.it

Abstract. We present the methodology and the experimental setting of the participation of the IMS Unipd team in TREC Clinical Trials 2021. The objective of this work is to continue the longitudinal study of the evaluation of query expansion, ranking fusion, and document filtering approach optimized in the previous participation to TREC.
In particular, we added to our procedure proposed in 2020, a comparison with a pipeline that use the large transformers.
The results obtained provide interesting insights in terms of the different per-topic effectiveness and will be used for further failure analyses.

Keywords: Precision medicine, query reformulation, rank fusion

1 Introduction

The TREC 2021 Clinical Trials (CT) Track[1] focuses on the problem of retrieving clinical trials given a lengthy query that describes the patient case that simulates an admission statement in an electronic health record.

Our participation to the TREC 2021 CT Track focuses on the evaluation of a mixture of query expansion, rank fusion, and document filtering approaches optimized on the experimental analyses of our previous participation to this track [5]. Therefore, the objective of this work is to continue the evaluation of this longitudinal study of different combinations of approaches. Moreover, for the first time, we added transformer-based [8] models in our pipeline of document analysis – namely, BART [4] and T5 [6].

In the following sections, we present the experiments we carried out using a fully automated system that:

- summarizes lengthy queries to reduce noise injection using transformer-based models;
- performs query expansion based on pseudo-relevance feedback information;
- filters out clinical trials for which a patient is not eligible based on age and gender information; and
- merges the different rankings produced by several approaches validated on previous TREC Precision Medicine collections.

[1] http://www.trec-cds.org/2021.html

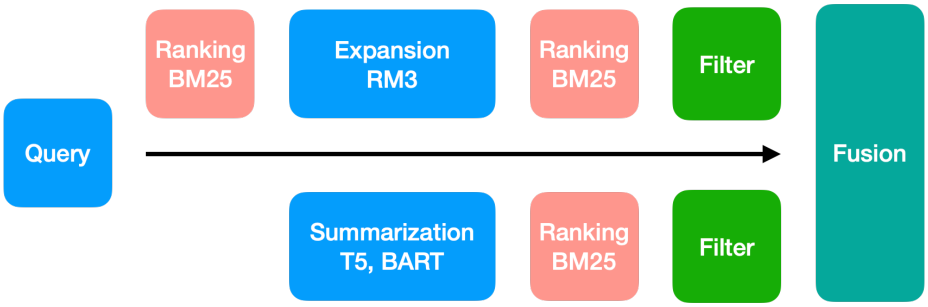


Fig. 1: Pipeline of the methodology: query expansion/summarization, filtering, and fusion

2 Methodology

In this section, we describe the methodology employed to conduct experiments. In particular, we merged the ranking lists provided by the different retrieval methods using (or not) summarized queries and applying query expansion based on pseudo-relevance feedback.

Query summarization: We use either BART [4] or T5 [6] models to perform summarization over the original, lengthy queries.

Query expansion: We used the RM3 model to implement a pseudo-relevance feedback strategy including query expansion [3, 2].

Retrieval models: For each query, we run the Okapi BM25 retrieval model [7].

Filtering: After the retrieval step, we filter out from the list of candidate trials those for which a patient is not eligible based on their demographic data – that is, age and gender. In other words, we automatically extract the patient's age and gender from queries and filter out trials with eligibility criteria that match the extracted age and gender values. In those cases where part of the demographic data are not specified, a clinical trial is kept or discarded on the basis of the remaining demographic information. For instance, if the clinical trial does not specify a required minimum age, then it is kept or discarded based on its maximum age and gender required values.

Ranking fusion: Given different ranking lists, we used the CombSUM [1] approach with minmax normalization to merge them.

measure	median	imsFused1	imsFused2	RM3Filtered	T5RM3Filt	BARTRM3Filt
NDCG@10	0.304	0.375	0.470	0.515	0.353	0.411
P@10	0.161	0.239	0.293	0.336	0.213	0.260
RecipRank	0.294	0.420	0.502	0.494	0.352	0.435

Table 1: Overall comparison with average median values of the scientific literature task

3 Experiments

For all the experiments, we used the PyTerrier search engine[2] with the following parameter settings for BM25:

- $k2 = 1.2$
- $b = 0.75$

3.1 Runs

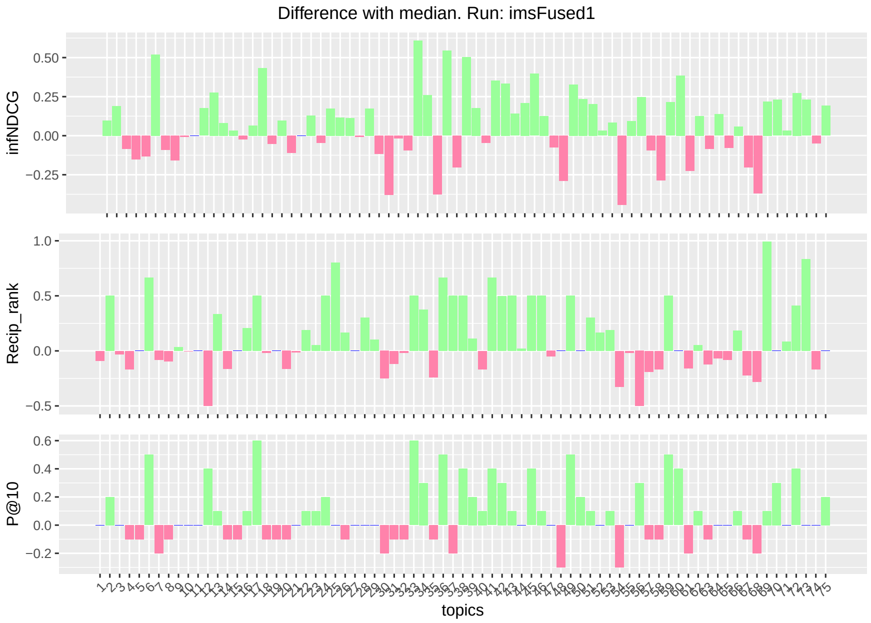
We submitted five runs:

- RM3Filtered: run with RM3 expansion, using BM25 as the first and second stage retrieval model. After both the first and the second retrieval stages, results have been filtered to remove trials with unfeasible age or sex attributes;
- T5RM3Filt: Prior to the retrieval, queries are summarized using the T5 summarization algorithm with a summary length - chosen by T5 - between 30 and 130 words. The same model as RM3Filtered is used to retrieve documents;
- BARTRM3Filt: Prior to the retrieval, queries are summarized using the BART summarization algorithm with a summary length - chosen by BART - between 30 and 130 words. The same model as RM3Filtered is used to retrieve documents;
- imsFused1: additive fusion of runs obtained with T5 summarizations with exact lengths 20, 50, 100, 150 and a run with T5 summarizations in the range 0-150. BM25 is used as the retrieval model. results with unfeasible values of age or sex have been removed;
- imsFused2: CombSUM fusion with min-max normalization of imsFused1, RM3Filtered, T5RM3Filt, and BARTRM3Filt;

3.2 Results

The organizers of the TREC 2021 PM Track provided the summary of the results in terms of best, median, and worst value for each topic for three evaluation measures: Normalized Discount Cumulative Gain (NDCG), precision at 10 (P@10), and Reciprocal Rank (RecipRank).

[2] `https://pyterrier.readthedocs.io/en/latest/`



(a)

Fig. 2: Topic by topic difference between the run and median values.

In Table 1, we report the median values of the three measures averaged across topics, as well as the averaged results of the five submitted runs.

In Figures 2a, 3a, 3b, 4a, 4b, we show a barplot that displays, topic by topic, the difference between the performance of each run and the median values of the task. For a positive difference (run better than median), a green barplot is shown, while for a negative difference (run worse than median), a red barplot is shown.

The results show that all the runs perform better than median values. In particular, the RM3 Filtered run performs significantly better than median (statistical analyses will be provided in the final version of the paper), followed by the imsFused2 run and the BART RM3 filtered rank. Given these promising results, we plan to investigate the integration of re-ranking components in the retrieval pipeline.

4 Final Remarks

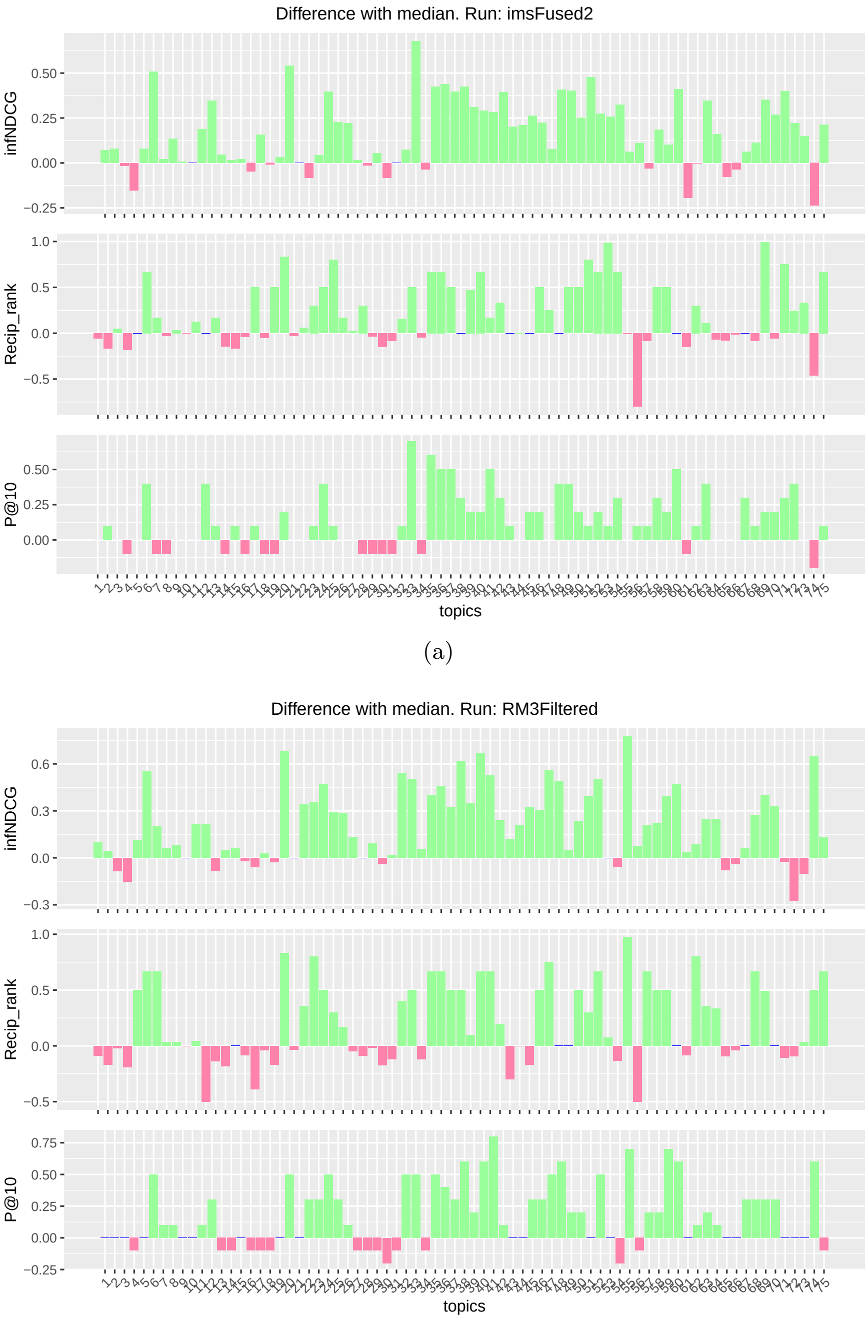
In this paper, we presented the results of our fourth participation in the TREC biomedical Track.

The analysis of the results showed the effectiveness of the filtered approach together with rank fusion runs for all the measures provided in the track.

As future work, we will investigate the possible optimization of the query summarization process in the ranking pipeline.

References

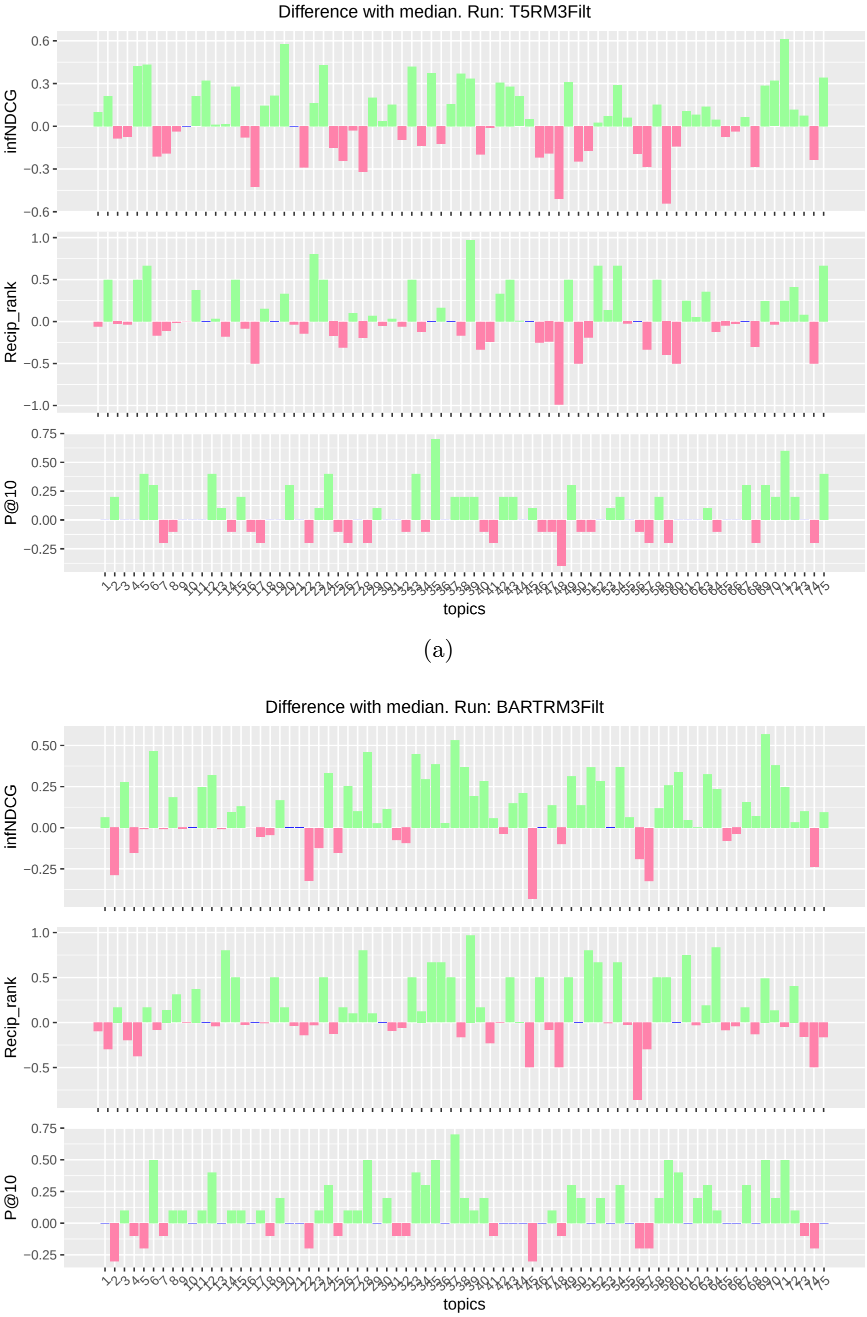
1. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. NIST special publication SP **243** (1994)
2. Jaleel, N.A., Allan, J., Croft, W.B., Diaz, F., Larkey, L.S., Li, X., Smucker, M.D., Wade, C.: Umass at TREC 2004: Novelty and HARD. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004. NIST Special Publication, vol. 500-261. National Institute of Standards and Technology (NIST) (2004), http://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf
3. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 120–127. SIGIR '01, Association for Computing Machinery, New York, NY, USA (2001). https://doi.org/10.1145/383952.383972, https://doi.org/10.1145/383952.383972
4. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 7871–7880. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.703, https://doi.org/10.18653/v1/2020.acl-main.703
5. Marchesin, S., Di Nunzio, G.M., Agosti, M.: Simple but effective knowledge-based query reformulations for precision medicine retrieval. Information **12**(10) (2021). https://doi.org/10.3390/info12100402, https://www.mdpi.com/2078-2489/12/10/402
6. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1–140:67 (2020), http://jmlr.org/papers/v21/20-074.html
7. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (2009). https://doi.org/10.1561/1500000019, https://doi.org/10.1561/1500000019
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html



(a)

(b)

Fig. 3: Topic by topic difference between the run and median values.



(a)

(b)

Fig. 4: Topic by topic difference between the run and median values.