Summarize and Expand Queries in Clinical Trials Retrieval The IIIA Unipd at TREC 2022 Clinical Trials

Giorgio Maria Di Nunzio, Guglielmo Faggioli, Stefano Marchesin

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

Abstract. We present the methodology and the experimental setting we, the Intelligent Interactive Information Access (IIIA)¹ UNIPD team, used in the TREC Clinical Trials 2022. This work continues the long streak of studies we carried out at TREC Precision Medicine to evaluate the effectiveness of query reformulation, pseudo-relevance feedback, and document filtering. Compared to the procedure proposed in 2021, we introduced the use of manual summarization and removed rank fusion. The obtained results provide interesting insights on the different pertopic effectiveness and will be used for further analyses.

Keywords: Precision medicine \cdot Query summarization \cdot Pseudo Relevance Feedback.

1 Introduction

The TREC 2022 Clinical Trials (CT) Track² 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 2022 CT Track focuses on the evaluation of (manual) query summarization, together with pseudo-relevance feedback [4] and document filtering approaches – optimized on the experimental analyses of our previous participation to this track [5, 1]. Therefore, the objective of this work is to continue the evaluation of this longitudinal study of different query reformulation strategies.

In the following sections, we present the experiments we carried out using a semi-automated system that: i) exploits different manual query summaries to reduce noise injection; ii) performs query expansion based on pseudo-relevance feedback; iii) reweighs clinical trials for which a patient is not eligible based on age and gender information.

¹ https://iiia.dei.unipd.it

² http://www.trec-cds.org/2022.html

2 Di Nunzio et al.

type	text						
original A 19-year-old male came to clinic with some sexual concern. He recently of							
	gaged in a relationship and is worried about the satisfaction of his girlfriend.						
	He has a "baby face" according to his girlfriend's statement and he is n						
	muscular as his classmates. On physical examination, there is some pubic h						
	and poorly developed secondary sexual characteristics. He is unable to detec						
	coffee smell during the examination, but the visual acuity is normal. Ultra-						
	sound reveals the testes volume of 1-2 ml. The hormonal evaluation showe						
	serum testosterone level of 65 ng/dL with low levels of GnRH.						
NLS	A 19-year-old male with some sexual concern. There is some pubic hair and						
	poorly developed secondary sexual characteristics. The hormonal evaluation						
	showed serum test osterone level of 65 $\rm ng/dL$ with low levels of GnRH.						
KS	male hormonal evaluation serum testosterone low levels GnRH						
Table 1: Examples of query variants for topic 1.							

2 Methodology

In this section, we describe the methodology employed to conduct experiments. In particular, we used two manual summarization approaches: i) Natural Language Summary (NLS), where we reduced the original query keeping the structure of the language; and ii) Keyword Summary (KS), where we kept only terms that are likely to be relevant. As an additional experiment, we also tried a two-step summarization where we further summarize NLS summaries using the transformer-based [8] T5 model [6]. After summarization, we applied both pseudo-relevance feedback and document filtering.

Query summarization: We used two manual summarization approaches: NLS and KS. We describe each approach below.

NLS. The summary is written by performing only ablation operations from the original query. The aim is to keep, when possible, the minimal amount of information such as 1) the generality of the patient (sex and age), 2) the initial examination (for example, "on physical examination"), 3) the results of the examination (for example, "the patient shows ...")

KS. The keyword-based query is built by removing text from the NLS variant. In particular, we keep the information about the sex of the patient (i.e., male, woman, girl, boy, ...), and all the terms with a high 'termhood' [3].

In Table 1, we show an example of query variants for topic 1.

Query expansion: We used the RM3 model to implement a pseudo-relevance feedback strategy including query expansion [4,2].

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

Filtering: After the retrieval step, we penalized the score of candidate trials for which a patient is not eligible based on their demographic data – that is, age and gender. In other words, we automatically extracted patient's age and gender from queries and penalized trials whose eligibility criteria do not allow for the extracted age and gender values. We decided to penalize, rather than remove, trials due to the lower importance that exclusion criteria have compared to inclusion ones. Put simply, it is more important that target patients match inclusion criteria rather than meet the exclusion ones. Therefore, removing trials for which patients meet the exclusion criteria might hinder retrieval performance.

3 Experiments

For all the experiments, we used the PyTerrier search engine³. We used the following parameter settings for BM25:

$$-k2 = 1.2$$

- b = 0.75

Regarding the penalization of those trials for which target patients meet exclusion criteria, we applied a 20% discount on their scores.

3.1 Runs

We submitted five runs:

- (1) ims_BM25Filtered_s: run with NLS summaries using BM25. After retrieval, results have been filtered to penalize trials meeting exclusion criteria;
- (2) ims_RM3Filtered_s: run with NLS summaries and 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 penalize trials meeting exclusion criteria;
- (3) ims_BM25Filtered_kw: run with KS summaries using BM25. After retrieval, results have been filtered to penalize trials meeting exclusion criteria;
- (4) ims_RM3Filtered_kw: run with KS summaries and 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 penalize trials meeting exclusion criteria;
- (5) ims_T5summarizer: run with NLS summaries using BM25. After the manual summarization, a further, automated summarization step is performed using T5. Results have been filtered to penalize trials meeting exclusion criteria;

3.2 Results

The organizers of the TREC 2022 PM Track provided the summary of the results in terms of best, median, and worst value for each topic for three evaluation

³ https://pyterrier.readthedocs.io/en/latest/

4 Di Nunzio et al.

measure	median	(1)	(2)	(3)	(4)	(5)
infNDCG	.392	.410	.446	.550	.542	.450
P@10	.258	.300	.200	.300	.400	.300
RecipRank	.411	.500	.333	.500	.500	.333

Table 2: Overall comparison with average median values.

measures: inferred Normalized Discount Cumulative Gain (infNDCG), precision at 10 (P@10), and Reciprocal Rank (RecipRank).

In Table 2, 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 1a, 1b, 2a, 2b, 3, 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 the runs have mixed performances compared with median values. Among the different approaches, those using keyword-based summaries seem to achieve higher performance. On the other hand, the impact of RM3 to expand queries is not clear, and might hinder the performance – as for the ims_RM3Filtered_s run. Given these mixed results, we plan to deepen the investigation on manual summarization to understand what is the proper tradeoff between NLS and KS summaries.

4 Final Remarks

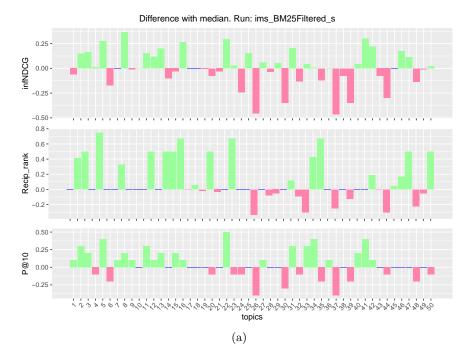
In this paper, we presented the results of our fifth participation in the TREC Medical Track. The analysis of the results showed mixed performance that needs to be further investigated to understand an effective tradeoff between different summarization strategies.

References

- Di Nunzio, G., Faggioli, G., Marchesin, S.: Filter, transform, expand, and fuse. In: Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, Gaithersburg, Maryland, USA, November 15-19, 2021. NIST Special Publication, National Institute of Standards and Technology (NIST) (2021)
- 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
- Kageura, K., Umino, B.: Methods of automatic term recognition : A review. Terminology 3, 259–289 (1996)

- 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
- 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
- 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-totext transformer. J. Mach. Learn. Res. 21, 140:1–140:67 (2020), http://jmlr.org/ papers/v21/20-074.html
- 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
- 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

6 Di Nunzio et al.





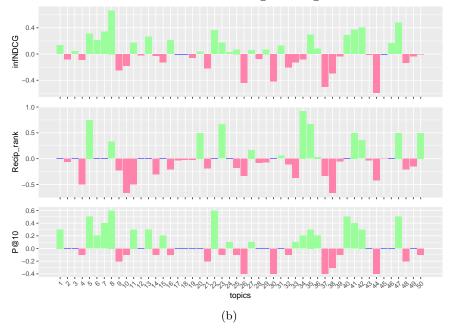
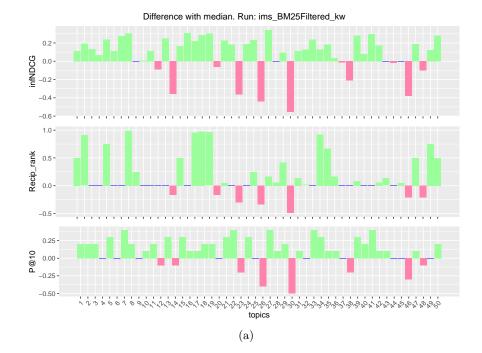


Fig. 1: Topic by topic difference between the run and median values.



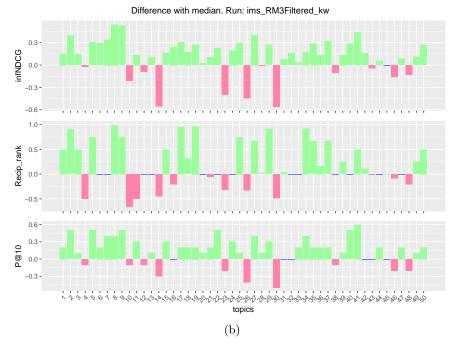


Fig. 2: Topic by topic difference between the run and median values.

8 Di Nunzio et al.



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