

An Analysis of a Methodology and Experimental Results for the Retrieval of Clinical Trials

Discussion Paper

Giorgio Maria Di Nunzio¹, Guglielmo Faggioli¹ and Stefano Marchesin¹

¹Department of Information Engineering, University of Padua, Italy

Abstract

In this paper, we present our previous and current work about the methodology and the experimental analysis of a query reformulation, pseudo-relevance feedback, and document filtering approach. In particular, we present a summary of two studies carried out in the context of the TREC Precision Medicine track. The two original papers are [1] and [2].

Keywords

Precision medicine, Query reformulation, Pseudo Relevance Feedback

1. Introduction

The Clinical Trial Task¹ at the Text REtrieval Conference (TREC)² is an information retrieval challenge in the medical domain where the query is the a synthetic patient descriptions and the corpus is a large set of clinical trial descriptions. This task is part of a long run challenge related to the evaluation of Clinical Decision Support systems³ started in 2014. These tasks have sought to provide benchmark datasets and evaluate information retrieval systems focused on many of the most important information access problems in biomedicine.

The dataset provided by the organizers of the task (in both 2021 and 2022) consists of a set of topics, a brief patient case description, and a set of documents, a snapshot of ClinicalTrials.gov. This collection consists of a snapshot of all the clinical trials available on ClinicalTrials.gov on April 27, 2021. The data is available as XML, with this specific snapshot containing 375,581 clinical trial descriptions.

In this extended abstract, we summarize the methodologies and the experimental results that we achieved in the last two editions of the Clinical Trial Task, 2021 and 2022, where the main


IIR2023: 13th Italian Information Retrieval Workshop, 8th - 9th June 2023, Pisa, Italy

✉ giorgiomaria.dinunzio@unipd.it (G. Di Nunzio); guglielmo.faggioli@unipd.it (G. Faggioli); stefano.marchesin@unipd.it (S. Marchesin)

🌐 <https://www.dei.unipd.it/~dinunzio> (G. Di Nunzio); <https://www.dei.unipd.it/~faggioli> (G. Faggioli); <https://www.dei.unipd.it/~marches1> (S. Marchesin)

🆔 0000-0001-9709-6392 (G. Di Nunzio); 0000-0002-5070-2049 (G. Faggioli); 0000-0003-0362-5893 (S. Marchesin)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹<https://www.trec-cds.org/2022.html>

²<https://trec.nist.gov>

³<https://www.trec-cds.org>

focus was the retrieval of clinical trials given a lengthy query that describes the patient case that simulates an admission statement in an electronic health record.

Our participation to these two editions focused on the evaluation of a mixture of query reformulation, rank fusion, and document filtering approaches optimized on the experimental analyses of our previous participations to this track [3]. In addition, we also performed experiments with pseudo-relevance feedback [4]. The objective of these studies is to continue the evaluation of this longitudinal study of different combinations of approaches. These results were reported in [1] and [2].

2. Methodology

The methodology we employed in both participations was mainly based on a query reformulation approach, the merging of the ranking lists provided by the different retrieval methods using (or not) summarized queries and applying query expansion based on pseudo-relevance feedback. In the following sections, we describe the approach for each element of the retrieval pipeline.

2.1. Query reformulation

In the 2021 edition, we use either BART [5] or T5 [6] models to perform summarization over the original, lengthy queries. In the 2022 edition, 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 [7] T5 model [6]. After summarization, we applied both pseudo-relevance feedback and document filtering.

2.2. Query expansion

We used the RM3 model to implement a pseudo-relevance feedback strategy including query expansion [4, 8].

2.3. Retrieval models:

For each query, we run the Okapi BM25 retrieval model [9].

2.4. 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 patient’s age and gender from queries and filter out trials whose eligibility criteria do not allow for 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.

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

2.5. Ranking fusion

Given different ranking lists, we used the CombSUM [10] approach with minmax normalization to merge them.

3. Results

The organizers of the TREC tasks usually provide 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). In the following sections, we summarize the details of each experiment and the results for each year.

3.1. Results of 2021

In 2021, 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;

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.

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.

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.

3.2. Results 2022

In 2022, 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;

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.

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. Conclusions

In this extended abstract, we have presented a brief overview of the experimental results obtained in the last two editions of the TREC Clinical Trial task. The results of the proposed approach were well above the median results of the participants in both editions. For some evaluation measures, our results were included in the top 5 performing systems.

References

- [1] G. Di Nunzio, G. Faggioli, S. Marchesin, 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.
- [2] G. Di Nunzio, G. Faggioli, S. Marchesin, Summarize and expand queries in clinical trials retrieval. the iiiaupnd at trec 2022 clinical trials, in: Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, Gaithersburg, Maryland, USA, November 15-19, 2022, NIST Special Publication, National Institute of Standards and Technology (NIST), 2022.
- [3] S. Marchesin, G. M. Di Nunzio, M. Agosti, Simple but effective knowledge-based query reformulations for precision medicine retrieval, *Information* 12 (2021). URL: <https://www.mdpi.com/2078-2489/12/10/402>. doi:10.3390/info12100402.
- [4] V. Lavrenko, W. B. Croft, Relevance based language models, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, Association for Computing Machinery, New York, NY, USA, 2001, pp. 120–127. URL: <https://doi.org/10.1145/383952.383972>. doi:10.1145/383952.383972.
- [5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, 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, Association for Computational Linguistics, 2020, pp. 7871–7880. URL: <https://doi.org/10.18653/v1/2020.acl-main.703>. doi:10.18653/v1/2020.acl-main.703.
- [6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* 21 (2020) 140:1–140:67. URL: <http://jmlr.org/papers/v21/20-074.html>.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, 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, 2017, pp. 5998–6008. URL: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [8] N. A. Jaleel, J. Allan, W. B. Croft, F. Diaz, L. S. Larkey, X. Li, M. D. Smucker, C. Wade, Umass at TREC 2004: Novelty and HARD, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, volume 500-261 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2004. URL: <http://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf>.
- [9] S. E. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, *Foundations and Trends in Information Retrieval* 3 (2009) 333–389. URL: <https://doi.org/10.1561/15000000019>. doi:10.1561/15000000019.
- [10] E. A. Fox, J. A. Shaw, Combination of Multiple Searches, NIST special publication SP 243 (1994).