A Post-Analysis of Query Reformulation Methods for Clinical Trials Retrieval
DISCUSSION PAPER

Maristella Agosti[1], Giorgio Maria Di Nunzio[1,2], and Stefano Marchesin[1]

[1]Department of Information Engineering, [2]Department of Mathematics
University of Padua, Italy
{maristella.agosti, giorgiomaria.dinunzio, stefano.marchesin}@unipd.it

Abstract. The Precision Medicine (PM) track of the Text REtrieval Conference (TREC) focuses on providing useful precision medicine information to clinicians treating cancer patients. The PM track gives the unique opportunity to evaluate medical IR systems on two different collections: scientific literature and clinical trials. In this paper, we evaluate several state-of-the-art query expansion and reduction methods to see whether a particular approach can be helpful in clinical trials retrieval. We present those approaches that are consistently effective in all three TREC PM editions and we compare them to the results obtained by the research groups who participated in all three editions.

Keywords: Query reformulation · knowledge base · precision medicine

1 Introduction and Motivations

Medical Information Retrieval (IR) helps a wide variety of users to access and search medical information archives and data [6]. In [9], a classification of textual medical information is proposed: 1) Patient-specific information which applies to individual patients. This type of information can be structured, as in the case of an Electronic Health Record (EHR), or can be free narrative text. 2) Knowledge-based information that has been derived and organized from observational or experimental research. In the case of clinical research, the information is most commonly provided by books and journals, but can take a wide variety of other forms. Therefore, the design of effective tools to access and search textual medical information requires, among other things, enhancing the query through expansion and/or rewriting methods that leverage the information contained within knowledge resources. [15] identified some challenges arising from the differences between general and medical case-based retrieval. In particular,

state-of-the-art retrieval methods, combined with selective query term weighing based on medical thesauri and physician feedback, improve performance significantly [16, 5]. In 2017, 2018, and 2019 the PM [11] track[1] at TREC[2] focused on an important use case in clinical decision support: providing useful precision medicine information to clinicians treating cancer patients. This track gives a unique opportunity to evaluate medical IR systems since the experimental collection is composed of a set of topics (synthetic cases created by precision oncologists) for two different collections that target two different tasks: 1) retrieving biomedical articles addressing relevant treatments for a given patient, and 2) retrieving clinical trials for which a patient – described in the information need – is eligible.

This work combines and discusses the methodology and the results originally presented at SIGIR 2019 [2] and TREC 2019 [4]. The objective is to evaluate several state-of-the-art query expansion and reduction methods to examine whether a particular approach can be helpful in clinical trials retrieval. Precisely, we compare the results obtained with our approach to the best experiments submitted to the 2017 and 2018 PM tracks [2]. Then, we select the top three query reformulations found in 2017 and 2018 PM tracks and we evaluate whether their effectiveness also holds in the 2019 PM track [4]. We conduct a systematic comparison between our approach and those proposed by the research groups that participated in all three years of TREC PM. The analysis shows the effectiveness of the proposed query reformulations in 2017 and 2018 PM tracks and confirms the trend in the 2019 PM track. The obtained runs achieve top performing results in all PM tracks [11–13]. In particular, a specific query reformulation allows the retrieval system to achieve top results in all three PM tracks.

The rest of the paper is organized as follows: Section 2 describes the approach used to evaluate different query reformulation methods. Section 3 presents the experimental setup and Section 4 compares the results obtained using our approach with those obtained by the other research groups who participated in TREC PM 2017, 2018 and 2019. Finally, Section 5 reports some final remarks.

2 Approach

The approach we propose consists of four steps: (i) indexing, (ii) query reformulation, (iii) retrieval and (iv) filtering.

Indexing. We create the following fields to index clinical trials collections: `<docid>`, `<text>`, `<max_age>`, `<min_age>` and `<gender>`. Fields `<max_age>`, `<min_age>` and `<gender>` contain information extracted from the `eligibility` section of clinical trials and are used in the filtering step. The `<text>` field contains the entire content of each clinical trial.

[1] `http://www.trec-cds.org/`
[2] `https://trec.nist.gov/`

Query Reformulation. The approach relies on two types of query reformulation methods: query expansion and query reduction.

Query expansion: We perform a knowledge-based a priori query expansion. First, we rely on MetaMap [3], a state-of-the-art medical concept extractor, to extract from each query field all the Unified Medical Language System (UMLS)[3] concepts belonging to the following semantic types:[4] `Neoplastic Process` (*neop*), `Gene or Genome` (*gngm*) and `Cell or Molecular Dysfunction` (*comd*). The *gngm* and *comd* semantic types are related to the query `<gene>` field, while *neop* is related to the `<disease>` field. For those collections where an additional `<other>` field is included – which considers other potential factors that may be relevant – MetaMap is used on `<other>` with no restriction on the semantic types, as its content does not refer to any particular semantic type. Second, for each extracted concept, we consider all its name variants contained into the following knowledge sources: National Cancer Institute (NCI), Medical Subject Headings (MeSH), SNOMED CT (SNOMEDCT) and UMLS Metathesaurus (MTH). All knowledge sources are manually curated and up-to-date. The expanded queries consist of the union of the original terms with the set of name variants. For example, consider a query only containing the word "*melanoma*" – which is mapped to the UMLS concept `C0025202`. The set of name variants for the concept "*melanoma*" contains, among many others: cutaneous melanoma, malignant melanoma, malignant melanoma (disorder). Therefore, the final expanded query is the union of the original term "*melanoma*" with all its name variants. Additionally, we expand queries that do not mention any kind of blood cancer (e.g. "lymphoma" or "leukemia") with the term *solid*. This expansion proved to be effective in [7] where the authors found that a large part of relevant clinical trials do not mention the exact disease. A more general term like *solid tumor* is preferable and more effective.

Query reduction: We reduce original queries by removing, whenever present, gene mutations from the `<gene>` field. To clarify: consider a topic where the `<gene>` field mentions "BRAF (V600E)". After reduction, the `<gene>` field becomes "BRAF". The reduction aims at mitigating the over-specificity of topics, since the information contained in a topic is too specific compared to those contained in the target documents [10]. Additionally, we remove the `<other>` field from those collections that include it – since it contains additional factors that are not necessarily relevant, thus representing a potential source of noise in retrieving precise information for patients.

Retrieval. We use BM25 [14] as retrieval model. Query terms obtained through query expansion are weighted lower than 1.0 to avoid introducing too much noise in the retrieval process [8].

Filtering. The `eligibility` section in clinical trials comprises three important demographic aspects that a patient needs to satisfy to be considered eligible for

[3] `https://www.nlm.nih.gov/research/umls/`
[4] `https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml`

the trial, namely: `minimum age`, `maximum age` and `gender`; where `minimum age` and `maximum age` are the minimum and the maximum age, respectively, required for a patient to be considered eligible for the trial, while `gender` is the required gender. After the retrieval step, we filter out from the list of candidate trials those for which a patient is not eligible – i.e. his/her demographic data (age and gender) does not satisfy the three aforementioned eligibility criteria. In those cases where part of the demographic data is 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.

3 Experimental Setup

This section describes the experimental collections and the setup used to apply and evaluate our approach.

Experimental Collections. We report the main information related to topics and document collections below.

Topics consist of 30, 50, and 40 synthetic cases created by precision oncologists in 2017, 2018, and 2019, respectively. In 2017, each topic contained four key elements in a semi-structured format: (1) disease (e.g. a type of cancer), (2) genetic variants (primarily present in tumors), (3) demographic information (e.g. age, gender), and (4) other factors (which could impact certain treatment options). In 2018 and 2019, each topic had three of the four key elements used in 2017: (1) disease, (2) genetic variants, and (3) demographic information. Furthermore, the 2019 topics contain ten non-cancer-related topics.

Clinical Trials consist of a set of 241,006 clinical trial descriptions, for both 2017 and 2018, and of an updated version of 306,238 descriptions for 2019. The collections are derived from ClinicalTrials.gov[5] – a database of privately and publicly funded clinical studies conducted around the world. When none of the available treatments are effective on oncology patients, the common recourse is to determine if any potential treatments are undergoing evaluation in a clinical trial. Therefore, it would be helpful to automatically identify the most relevant clinical trials for an individual patient. Precision oncology trials typically use a certain treatment for a certain disease with a specific genetic variant. Such trials can have complex inclusion and/or exclusion criteria that are challenging to match with automated systems.

Experimental Procedure. We use Whoosh,[6] a Python search engine library, for indexing, retrieval, and filtering. For BM25, we keep the default values $k_1 = 1.2$ and $b = 0.75$ provided – as we found them to be a good combination [1]. For query expansion, we rely on MetaMap to extract and disambiguate concepts from UMLS. Below we report the procedure used for each experiment.

[5] `https://clinicaltrials.gov/`
[6] `https://whoosh.readthedocs.io/en/latest/intro.html`

- *Indexing*
  - Index clinical trials using the following created fields: `<docid>`, `<text>`, `<max_age>`, `<min_age>` and `<gender>`.

- *Query reformulation*
  - Use MetaMap to extract from each query field the UMLS concepts restricted to the following semantic types: *neop* for `<disease>`, *gngm*/*comd* for `<gene>`, and *all* for `<other>`;
  - Extract from concepts all name variants belonging to NCI, MeSH, SNOMED CT and MTH knowledge sources;
  - Expand (or not) topics that do not mention "lymphoma" or "leukemia" with the term *solid*;
  - Reduce (or not) queries by removing, whenever present, gene mutations from the `<gene>` field;
  - Remove (or not) the `<other>` field.

- *Retrieval*
  - 2017/2018 PM track: Adopt any combination of the reformulation strategies;
  - 2019 PM track: Adopt the three best reformulation strategies from 2017/2018 PM tracks;
  - Weigh expanded terms with a value $k = 0.1$;
  - Perform a search using expanded queries with BM25.

- *Filtering*
  - Filter out clinical trials for which the patient is not eligible.

Evaluation Measures. We use the measures adopted in the TREC PM tracks, that are: inferred nDCG (infNDCG), R-precision (Rprec) and P@10.

4 Experimental Results and Discussion

In Table 1, we report the results of our experiments on query reformulation (Part A) and compare them with the results obtained by the research groups that participated at TREC PM 2017, 2018 and 2019 (Part B). Given the large number of experiments we performed, we decided to only present the 5 runs with the highest P@10 for each year. Each line shows a particular combination (*y*es or *n*o values) of semantic types (*neop*, *comd*, *gngm*), usage and expansion of `<other>` field (oth, oth_exp), query reduction (orig), and expansion using weighted *solid* (tumor) keyword. We use the symbol '·' to indicate that the features oth, oth_exp are not applicable for the years 2018 and 2019 due to the absence of the `<other>` field in 2018 and 2019 topics. We highlight in bold the top 3 scores for each measure, and we use the symbol ‡ to indicate the combination that performs well in all three years. For the TREC PM participants, we select those participants who submitted runs in all three years and reached the top 10 performing runs in at least one edition for each measure [11–13]. The results

A: Analysis of Query Reformulations

line	year	neop	comd	gngm	oth	oth_exp	orig	solid	P@10	infNDCG	Rprec
1	2017	n	n	y	n	n	n	0.1	**0.3931**	-	**0.3263**
2‡	2017	n	n	n	n	n	n	0.1	**0.4034**	-	**0.3361**
3	2017	y	n	n	n	n	n	0.1	0.3862	-	0.3202
4	2017	n	n	n	n	n	n	n	**0.3931**	-	0.3241
5	2017	n	n	y	n	n	y	n	0.3862	-	**0.3243**
6	2018	n	n	n	·	·	y	n	0.5680	**0.5411**	**0.4197**
7	2018	n	n	n	·	·	y	0.1	**0.5740**	**0.5403**	**0.4179**
8	2018	y	n	n	·	·	n	n	**0.5700**	0.5345	0.4134
9‡	2018	n	n	n	·	·	n	0.1	**0.5820**	**0.5446**	**0.4205**
10	2018	y	n	y	·	·	n	n	0.5680	0.5393	0.4122
11	2019	n	n	n	·	·	y	0.1	**0.5368**	**0.6239**	**0.4386**
12	2019	y	n	n	·	·	n	n	0.5237	0.5755	0.4135
13‡	2019	n	n	n	·	·	n	0.1	**0.5316**	**0.5940**	**0.4264**
14	2019	n	n	y	·	·	n	0.1	**0.5263**	**0.6070**	**0.4302**
15	2019	n	n	n	·	·	n	n	0.5105	0.5853	0.4239

B: Comparison with TREC PM other Participants

line	year	TREC PM Participant Identifier	P@10	infNDCG	Rprec
1	2017	BiTeM	0.3586	-	-
2	2017	cbnu	<	-	-
3	2017	CSIROmed	<	-	-
4	2017	ECNUica	<	-	-
5	2017	POZNAN_SEMMED	0.3690	-	-
	2017	Top 10 threshold	0.3586	-	-
	2017	Best combination of our approach (A.2‡)	0.4034	-	0.3361
6	2018	SIBTextMining	<	<	<
7	2018	cbnu	<	<	<
8	2018	CSIROmed	<	<	<
9	2018	ECNUica	<	<	<
10	2018	Poznan	0.5580	0.4894	0.4101
	2018	Top 10 threshold	0.5240	0.4736	0.3658
	2018	Best combination of our approach (A.9‡)	0.5820	0.5446	0.4205
11	2019	BITEM_PM	0.4711	0.4963	0.3698
12	2019	cbnu	0.4921	0.5568	0.4121
13	2019	CSIROmed	0.4921	0.4930	0.3586
14	2019	ECNU-ICA	0.5053	0.5355	0.4001
15	2019	POZNAN	0.4421	0.4810	0.3503
	2019	Top 10 threshold	0.3658	0.4320	0.3230
	2019	Best combination of our approach (A.11)	0.5368	0.6239	0.4386

Table 1: Results for the TREC PM tasks. Part A (top) reports the results achieved using the five most effective query reformulations for each year. (‡) indicates a particular query reformulation effective in all three years. Part B (bottom) reports the results obtained by participant runs, along with the lowest score required to enter the top 10 TREC results list and the score obtained by the best combination of our approach. Further details are reported in Section 4.

reported in part B of Table 1 indicate the best score obtained by a particular run for a specific measure; the best results of a participant are often related to different runs. The symbol '−' means that the measure is not available, while '<' indicates that none of the runs submitted by the participant achieved the top 10 performing runs. For the sake of comparison, we add for each measure the lowest score required to enter the top 10 TREC results list, and the score obtained by the best combination of our approach – indicated by the line number – as if we were participants of these tracks.

Analysis of Query Reformulations. The results from Table 1 (Part A) highlight that the use of *solid* expansions with weight 0.1 as well as query gene reductions ($orig = n$) seems to improve performance consistently in 2017 – two of the three best runs in terms of P@10 (lines 1 and 2) applying both techniques. Regarding knowledge-based expansions, the semantic type *gngm* (lines 1 and 5) seems more effective than *neop* (line 3), while *comd* does not seem to have any positive effect at all. All five runs do not consider the other field ($oth = n$) nor its expansion ($oth\_exp = n$) – confirming our intuition that it might represent a potential source of noise in retrieving precise information for patients. Similarly to 2017, two of the best three runs of 2018 use no knowledge-based expansions and rely on the *solid* (tumor) expansion with weight 0.1 (lines 7 and 9). In particular, the runs combining query gene reductions and *solid* expansions (marked as ‡) provide the best performances for all the measures considered, both in 2017 and 2018. This suggests that removing highly specialized information (i.e. the gene mutation) or adding general terms (e.g. solid) benefits the retrieval. A possible reason is related to the nature of the document collections, since clinical trials often contain general requirements to allow patients to enroll. The results obtained in 2019 with the top three query reformulations from 2017 and 2018 confirm this trend. The run combining query gene reductions and *solid* expansions (line 13‡) is one of the top three runs of 2019, however two query reformulations from 2017 (line 14) and 2018 (line 11) provide better performance. This result shows how difficult the task is. In fact, even though we found a particular query reformulation approach (marked as ‡) to be highly effective in all three years – especially in 2017 and 2018 – it was not the best approach for 2019. Therefore, this analysis helps researchers to select an effective subset of query reformulations to build strong baselines for clinical trials retrieval.

Comparison with TREC PM Participants. The results from Table 1 (Part B) mark a clear division between the 2017 and 2018 tracks and the 2019 track. In 2017 and 2018, most of the participant runs did not reach the top 10 threshold in any of the considered measures – the only exception is the research group from Poznan University of Technology, whose best runs always belong to the top 10 performing runs for the track. Conversely, in 2019 all the participant groups submitted runs that achieved results higher than the top 10 threshold.

Compared with the results obtained using the query reformulations from Table 1 (Part A), we can see that all runs employing the best query reformulations obtain results higher than the top 10 threshold for all the considered measures in all three years. Furthermore, the runs using the top five query reformulations achieved consistently better results than participant runs for each measure in all three years. This is an indication of the robustness of our approach across the different collections and also of the effectiveness of the proposed query reformulations for the clinical trials retrieval. In particular, it is worth to mention that the runs using the (‡) query reformulation achieve performances that belong to the top three best performing systems of each year PM track [11–13]. Therefore, the analysis of query reformulations made on the 2017 and 2018 PM tracks con-

firmed its trend in the 2019 PM track and allowed us to identify a specific set of query reformulations beneficial for the retrieval of clinical trials.

5 Conclusions and Final Remarks

In this paper, we further elaborate the results originally presented at SIGIR 2019 [2] and TREC 2019 [4] to evaluate several query expansion and reduction methods and to see whether a particular approach can be helpful in clinical trials retrieval. The experimental analysis showed the effectiveness of the proposed query reformulations in 2017 and 2018 PM tracks, and we confirmed this positive trend in the 2019 PM track. The obtained runs achieve top performing results in all PM tracks [11–13]. In particular, the run marked as ‡ in Table 1 can be considered as a valid baseline to build stronger multi-stage systems in the future.

Acknowledgements. This work is partially supported by the ExaMode project, as part of the European Union H2020 research and innovation program under grant agreement no. 825292.

References

1. Agosti, M., Di Nunzio, G.M., Marchesin, S.: The University of Padua IMS Research Group at TREC 2018 PM Track. In: Proc. TREC (2018)
2. Agosti, M., Di Nunzio, G.M., Marchesin, S.: An Analysis of Query Reformulation Techniques for Precision Medicine. In: Proc. ACM SIGIR Conf. pp. 973–976 (2019)
3. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proc. AMIA Symposium. pp. 17–21 (2001)
4. Di Nunzio, G.M., Marchesin, S., Agosti, M.: Exploring how to combine query reformulations for precision medicine. In: Proc. TREC (2019)
5. Diao, L., et alii: The Research of Query Expansion Based on Medical Terms Reweighting in Medical IR. EURASIP J. Wireless Comm.&Networ. (1), 105 (2018)
6. Goeuriot, L., Jones, G., Kelly, L., Müller, H., Zobel, J.: Medical Information Retrieval: Introduction to the Special Issue. Inform Retrieval J. **19**(1), 1–5 (2016)
7. Goodwin, T.R., Skinner, M.A., Harabagiu, S.M.: UTD HLTRI at TREC 2017: Precision medicine track. In: Proc. TREC (2017)
8. Gurulingappa, H., Toldo, L., Schepers, C., Bauer, A., Megaro, G.: Semi-supervised information retrieval system for clinical decision support. In: Proc. TREC (2016)
9. Hersh, W.: Information Retrieval: A Health and Biomedical Perspective. Health and Informatics Series, Springer-Verlag, New York, NY, USA (2009)
10. Oleynik, M., et alii: HPI-DHC at TREC 2018: PM Track. In: Proc. TREC (2018)
11. Roberts, K., et alii: Overview of PM Track. In: Proc. TREC (2017)
12. Roberts, K., et alii: Overview of PM Track. In: Proc. TREC (2018)
13. Roberts, K., et alii: Overview of PM Track. In: Proc. TREC (2019)
14. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2009)
15. Sondhi, P., et alii: Leveraging Medical Thesauri and Physician FB for Improving Medical Literature Retrieval for Case Queries. JAMIA **19**(5), 851–858 (2012)
16. Zhu, D., Wu, S., Carterette, B., Liu, H.: Using Large Clinical Corpora for QE in Text-Based Cohort Identification. J. of Biomedical Informatics **49**, 275–281 (2014)