An Analysis of Query Reformulation Techniques for Precision Medicine

Maristella Agosti, Giorgio Maria Di Nunzio, Stefano Marchesin
Department of Information Engineering
University of Padua, Italy
{maristella.agosti,giorgiomaria.dinunzio,stefano.marchesin}@unipd.it

ABSTRACT

The Precision Medicine (PM) track at the Text REtrieval Conference (TREC) focuses on providing useful precision medicine-related information to clinicians treating cancer patients. The PM track gives the unique opportunity to evaluate medical IR systems using the same set of topics on two different collections: scientific literature and clinical trials. In the paper, we take advantage of this opportunity and we propose and evaluate state-of-the-art query expansion and reduction techniques to identify whether a particular approach can be helpful in both scientific literature and clinical trial retrieval. We present those approaches that are consistently effective in both TREC editions and we compare the results obtained with the best performing runs submitted to TREC PM 2017 and 2018.

CCS CONCEPTS

• **Information systems → Specialized information retrieval**; **Ontologies**; **Query reformulation**.

KEYWORDS

Medical IR; query reformulation; precision medicine

ACM Reference Format:
Maristella Agosti, Giorgio Maria Di Nunzio, Stefano Marchesin. 2019. An Analysis of Query Reformulation Techniques for Precision Medicine. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331289

1 MOTIVATIONS

Medical Information Retrieval (IR) helps a wide variety of users to access and search medical information archives and data [4]. In [7, chapter 2], 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, including computerized media. 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 techniques that leverage the information contained within knowledge resources. In this context, Sondhi et al. [12] identified some challenges arising from the differences between general retrieval 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 [3, 13].

In 2017 and 2018, the Precision Medicine (PM) [10] track[1] at the Text REtrieval Conference (TREC)[2] focused on an important use case in clinical decision support: providing useful precision medicine-related 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.

The objective of our study is to take advantage of this opportunity and evaluate several state-of-the-art query expansion and reduction techniques to examine whether a particular approach can be helpful in both scientific literature and clinical trials retrieval. Given the large number of participating research groups to this TREC track, we are able to compare the best experiments submitted to the PM track based on the results which were obtained applying our approach in the last two years. The experimental analysis shows that there are some common patterns in query reformulation that allow the retrieval system to achieve top performing results in both tasks.

The rest of the paper is organized as follows: Section 2 describes the approach used to evaluate different query reformulation techniques. Section 3 presents the experimental setup and compares the results obtained using our approach with the best performing runs from TREC PM 2017 and 2018. Finally, Section 4 reports some final remarks and concludes the paper.

2 APPROACH

The approach we propose for query expansion/reduction in a PM task comprises three steps, plus an additional fourth step required only for the retrieval of clinical trials. The steps are: (i) indexing, (ii) query reformulation, (iii) retrieval and (iv) filtering.

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

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331289

Indexing Step. 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 required for the filtering step. The <text> field contains the entire content of each clinical trial — and therefore also the information stored within the fields described above.

To index scientific literature collections, we create the following fields: <docid> and <text>. As for clinical trials, the <text> field contains the entire content of each target document.

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

Query expansion: We perform a knowledge-based a priori query expansion. First, we rely on MetaMap [2], 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 consistently 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[5] (NCI), Medical Subject Headings[6] (MeSH), SNOMED CT[7] (SNOMEDCT) and UMLS Metathesaurus[8] (MTH). All knowledge sources are manually curated and up-to-date.

The expanded queries consist in 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); etc. 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 [5] 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 the reduction process, the <gene> field becomes "BRAF". The reduction process aims to mitigate the over-specificity of topics, since the information contained in a topic is too specific compared to those contained in the target documents [8].

[3]https://www.nlm.nih.gov/research/umls/
[4]https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml
[5]https://www.cancer.gov/
[6]https://www.ncbi.nlm.nih.gov/mesh/
[7]http://www.snomed.org/
[8]https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

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.[9]

Retrieval Step. We use BM25 [11] as retrieval model. Additionally, query terms obtained through query expansion are weighted lower than 1.0 to avoid introducing too much noise in the retrieval process [6].

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

Therefore, 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 aforementioned. 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 SETUP AND EVALUATION

In this section, we describe the experimental collections and the setup used to conduct and evaluate our approach. Then, we compare the results obtained with our approach with those of the best performing systems from TREC PM 2017 and 2018. All these systems make use of external knowledge sources to enhance retrieval performance; moreover, most of them are complex multi-stage retrieval systems, like those proposed in [5, 8], while the approach we present is quite simple and straightforward – facilitating its reproducibility.[10]

Experimental Collections. Both tasks in TREC PM use the same set of topics, but with two different collections: scientific literature, clinical trials.

Topics consists of 30 and 50 synthetic cases created by precision oncologists in 2017 and 2018, respectively. In 2017, topics contain 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, topics contain three of the four key elements used in 2017: (1) disease, (2) genetic variants, and (3) demographic information.

Scientific Literature consists of a set of 26,759,399 MEDLINE[11] abstracts, plus two additional sets of abstracts: (i) 37,007 abstracts from recent proceedings of the American Society of Clinical Oncology (ASCO), and (ii) 33,018 abstracts from recent proceedings of the American Association for Cancer Research (AACR). These

[9]In a personal communication with the organizers of the track, we have been informed that it was difficult to convince the oncologists why the *other* field was even necessary.
[10]Source code available at: https://github.com/stefano-marchesin/TREC_PM_qreforms
[11]https://www.nlm.nih.gov/bsd/pmresources.html

additional datasets were added to increase the set of potentially relevant treatment information. In fact, precision medicine is a fast-moving field where keeping up-to-date with the latest literature can be challenging due to both the volume and velocity of scientific advances. Therefore, when treating patients, it may be helpful to present the most relevant scientific articles for an individual patient. Relevant literature articles can guide precision oncologists to the best-known treatment options for the patient's condition.

Clinical Trials consists of a total of 241,006 clinical trial descriptions, derived from ClinicalTrials.gov[12] — a repository of clinical trials in the U.S. and abroad. 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 (or set of variants). Such trials can have complex inclusion and/or exclusion criteria that are challenging to match with automated systems.

Experimental Setup. We use Whoosh,[13] a pure Python search engine library, for indexing, retrieval and filtering steps. For BM25, we keep the default values $k_1 = 1.2$ and $b = 0.75$ provided by Whoosh – as we found them to be a good combination [1]. For query expansion, we rely on MetaMap to extract and disambiguate concepts from UMLS. We summarize the procedure used for each experiment below.

Indexing

- Index clinical trials using the following created fields: <docid>, <text>, <max_age>, <min_age> and <gender>;
- Index scientific abstracts using the following created fields: <docid> and <text>.

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 the 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

- Adopt any combination of the reformulation strategies;
- Weigh expanded terms with a value $k \in \{0, 0.1, 0.2, ..., 1\}$;
- 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 official measures adopted in the TREC PM track: inferred nDCG (infNDCG), R-precision (Rprec) and Precision at rank 10 (P_10). Precision at rank 5 and at rank 10 were used only for the Clinical Trials task 2017 and are not reported in this work for space reasons. The inferred nDCG was not computed for the task Clinical Trials 2017 since the sampled relevance judgments are not available.

Comparison. In Table 1, we report the results of our experiments (upper part) and compare them with the top performing participants at TREC 2017 and 2018 (lower part). Given the large number of experiments, we decided to present the top 5 runs ordered by P_10 for each year and for each task. 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 year 2018 due to the absence of the <other> field in 2018 topics. We report the results for both Scientific Literature (sl) and Clinical Trials (ct) tasks. We highlight in bold the top 3 scores for each measure, and we use the symbols † and ‡ to indicate two combinations that performed well in both 2017 and 2018. For the TREC PM participants, we select those participants who submitted runs in both years and reached the top 10 performing runs in at least two measures [9, 10]. The results reported in the lower part 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 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.

In 2018, there is a clear distinction in terms of performances among the combinations that achieve the best results for the *sl* and the *ct* tasks. For the *sl* task, considering the semantic type *neop* expansion without using the umbrella term *solid* provides the best performances for all the measures considered. On the other hand, two of the best three runs for the *ct* task (line 5 and 9), use no semantic type expansion, but rely on the *solid* (tumor) expansion with weight 0.1.

In 2017, the situation is completely different. Lines 12 and 13 show two combinations that are in the top 3 performing runs for both *sl* and *ct*. These two runs use query reduction and a weighted 0.1 *solid* (tumor) expansion. The use of a weighted 0.1 *solid* expansion as well as a reduced query (*orig* = *n*) seems to improve performances consistently for all measures in 2017. The semantic type *gngm* seems more effective than *neop*, while *comd* does not seem to have any positive effect at all.

Another element that shows how difficult these two tasks are is the fact that top performing systems in 2017 do not achieve the same results in 2018. Our study therefore helps researchers to select (or remove) semantic types to build strong baselines for both tasks.

4 CONCLUSIONS AND FINAL REMARKS

In this paper, we proposed and evaluated several state-of-the-art query expansion and reduction techniques for scientific literature and clinical trials retrieval. The experimental analysis showed that no clear pattern emerges for both tasks. In general, a query expansion approach using a selected set of semantic types helps the

[12] https://clinicaltrials.gov/
[13] https://whoosh.readthedocs.io/en/latest/intro.html

		Semantic Type			Field Other				sl	ct	sl	ct	sl	ct
line	year	neop	comd	gngm	oth	oth_exp	orig	solid	P_10	P_10	infNDCG	infNDCG	Rprec	Rprec
1	2018	y	y	n	·	·	y	n	**0.5660**	0.5540	**0.4912**	0.5266	**0.3288**	0.4098
2	2018	y	n	n	·	·	y	n	**0.5640**	0.5600	**0.4961**	0.5264	**0.3288**	0.4138
3	2018	y	n	y	·	·	y	n	**0.5480**	0.5660	**0.4941**	0.5292	**0.3266**	0.4116
4	2018	n	n	n	·	·	y	n	0.5460	0.5680	0.4876	**0.5411**	0.3240	**0.4197**
5	2018	n	n	n	·	·	y	0.1	0.5440	**0.5740**	0.4877	**0.5403**	0.3247	**0.4179**
6	2018	n	y	n	·	·	y	n	0.5440	0.5540	0.4853	**0.5403**	0.3236	0.4130
7	2018	y	n	n	·	·	n	n	0.5420	**0.5700**	0.4636	0.5345	0.3180	0.4134
8†	2018	n	n	y	·	·	y	n	0.5340	0.5640	0.4877	0.5337	0.3229	0.4106
9‡	2018	n	n	n	·	·	n	0.1	0.5300	**0.5820**	0.4635	**0.5446**	0.3148	**0.4205**
10	2018	y	n	y	·	·	n	n	0.5140	0.5680	0.4572	0.5393	0.3144	0.4122
11	2017	y	n	y	n	n	n	0.1	**0.5033**	0.3759	**0.3984**	-	0.2697	0.3206
12	2017	n	n	y	n	n	n	0.1	**0.4900**	**0.3931**	0.3881	-	0.2677	**0.3263**
13‡	2017	n	n	n	n	n	n	0.1	**0.4800**	**0.4034**	0.3931	-	**0.2728**	**0.3361**
14	2017	y	n	n	n	n	n	0.1	0.4767	0.3862	**0.3974**	-	**0.2714**	0.3202
15	2017	n	n	n	n	n	n	n	0.4733	**0.3931**	**0.3943**	-	**0.2732**	0.3241
16	2017	y	n	y	n	n	y	0.1	0.4733	0.3828	0.3567	-	0.2329	**0.3253**
17†	2017	n	n	y	n	n	y	n	0.4633	0.3862	0.3442	-	0.2254	0.3243
TREC PM Participant Identifier														
18	2018	UTDHLTRI							0.6160	0.5380	0.4797	0.4794	<	0.3920
19	2018	UCAS							0.5980	0.5460	0.5580	0.5347	0.3654	0.4005
20	2018	udel_fang							0.5800	0.5240	0.5081	0.5057	0.3289	0.3967
21	2018	NOVASearch							<	0.5520	<	0.4992	<	0.3931
22	2018	Poznan							<	0.5580	<	0.4894	<	0.4101
	2018	Top 10 threshold							0.5800	0.5240	0.4710	0.4736	0.2992	0.3658
	2018	Best combination of our approach							(1) 0.5660	(9‡) 0.5820	(2) 0.4961	(9‡) 0.5446	(1) 0.3288	(9‡) 0.4205
23	2017	UTDHLTRI							0.6300	0.4172	0.4647	-	0.2993	-
24	2017	udel_fang							0.5067	<	0.3897	-	0.2503	-
25	2017	NOVASearch							<	0.3966	<	-	<	-
26	2017	Poznan							<	0.3690	<	-	<	-
27	2017	UCAS							<	0.3724	<	-	0.2282	-
	2017	Top 10 threshold							0.4667	0.3586	0.3555	-	0.2282	-
	2017	Best combination of our approach							(11) 0.5033	(13‡) 0.4034	(11) 0.3984	-	(15) 0.2732	(13‡) 0.3361

Table 1: Results for the TREC PM tasks 2017 and 2018. Details are reported in Section 3.

retrieval of scientific literature, while a query reduction approach without expansion, but a small weighted solid (tumor) keyword expansion, improves performances of the clinical trials task. Nevertheless, we found that a particular combination (marked as ‡) performs well in both tasks – in particular the clinical trials task – and could have been one of the top 10 performing runs across many evaluation measures in both TREC PM 2017 and 2018. Therefore, this run can be considered as a baseline on which stronger multi-stage systems can be built.

ACKNOWLEDGMENTS

The authors thank Ellen Vorhees and Kirk Roberts for their helpful insights regarding the interpretation of the data collection. The work was partially supported by the ExaMode project,[14] as part of the European Union H2020 research and innovation program under grant agreement no. 825292.

REFERENCES

[1] M. Agosti, G.M. Di Nunzio, and S. Marchesin. 2018. The University of Padua IMS Research Group at TREC 2018 Precision Medicine Track. In *Proc. of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, Nov. 14-16, 2018.*

[2] A.R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proc. of the AMIA Symposium.* American Medical Informatics Association, 17–21.

[3] L. Diao, H. Yan, F. Li, S. Song, G. Lei, and F. Wang. 2018. The Research of Query Expansion Based on Medical Terms Reweighting in Medical Information Retrieval. *EURASIP Jour. on Wireless Communications and Networking* 2018, 1 (04 May 2018), 105. https://doi.org/10.1186/s13638-018-1124-3

[4] L. Goeuriot, G.J.F. Jones, L. Kelly, H. Müller, and J. Zobel. 2016. Medical Information Retrieval: Introduction to the Special Issue. *Information Retrieval Journal* 19, 1 (01 Apr 2016), 1–5. https://doi.org/10.1007/s10791-015-9277-8

[5] T.R. Goodwin, M.A. Skinner, and S.M. Harabagiu. 2017. UTD HLTRI at TREC 2017: Precision Medicine Track. In *Proc. of the Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, Nov. 15-17, 2017.*

[6] H. Gurulingappa, L. Toldo, C. Schepers, A. Bauer, and G. Megaro. 2016. Semi-Supervised Information Retrieval System for Clinical Decision Support. In *Proc. of the Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, Nov. 15-18, 2016.*

[7] W. Hersh. 2009. *Information Retrieval: A Health and Biomedical Perspective.* 2009 Springer Science + Business Media, LLC, New York, NY, USA.

[8] M. Oleynik, E. Faessler, A. Morassi Sasso, A. Kappattanavar, B. Bergner, H. Freitas da Cruz, J.P. Sachs, S. Datta, and E. Böttinger. 2018. HPI-DHC at TREC 2018: Precision Medicine Track. In *Proc. of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, Nov. 14-16, 2018.*

[9] K. Roberts, D. Demner-Fushman, E.M. Voorhees, W.R. Hersh, S. Bedrick, and A.J. Lazar. 2018. Overview of the TREC 2018 Precision Medicine Track. In *Proc. of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, Nov. 14-16, 2018.*

[10] K. Roberts, D. Demner-Fushman, E.M. Voorhees, W.R. Hersh, S. Bedrick, A.J. Lazar, and S. Pant. 2017. Overview of the TREC 2017 Precision Medicine Track. In *Proc. of the Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, Nov. 15-17, 2017.*

[11] S. Robertson and H. Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[12] P. Sondhi, J. Sun, C. Zhai, R. Sorrentino, and M.S. Kohn. 2012. Leveraging Medical Thesauri and Physician Feedback for Improving Medical Literature Retrieval for Case Queries. *Jour. of the American Medical Informatics Association: JAMIA* 19, 5 (Sep-Oct 2012), 851–858. https://doi.org/10.1136/amiajnl-2011-000293

[13] D. Zhu, S. Wu, B. Carterette, and H. Liu. 2014. Using Large Clinical Corpora for Query Expansion in Text-Based Cohort Identification. *Jour. of Biomedical Informatics* 49 (2014), 275 – 281. https://doi.org/10.1016/j.jbi.2014.03.010

[14] http://www.examode.eu/