Exploring how to Combine Query Reformulations for Precision Medicine

Giorgio Maria Di Nunzio, Stefano Marchesin, Maristella Agosti

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

Abstract. We report on our participation as the IMS Unipd team in both TREC PM 2019 tasks. The objective of the work is twofold: (i) we want to evaluate how different query reformulations affect the results and whether the findings obtained in previous years remain valid; (ii) we want to verify if combining different query reformulations based on expansion and reduction techniques prove effective in such a highly specific scenario. In particular, we designed a procedure to (1) filter out clinical trials based on demographic data, (2) perform query reformulations – both expansion and reduction techniques – based on knowledge bases to increase the probability of findings relevant documents, (3) apply rank fusion techniques to the rankings produced by the different query reformulations. We consider those query reformulations that have been validated on previous TREC PM experimental collections. These queries represent the most effective reformulations for our system on those topics/collections. The results obtained – especially in the clinical trials task – validate our assumptions and provide interesting insights in terms of the different per-topic effectiveness of the query reformulations.

Keywords: Precision medicine, query reformulation, rank fusion

1 Introduction

The TREC 2019 Precision Medicine (PM) Track[1] focuses on a relevant use case in clinical decision support: to provide useful precision medicine-related information to clinicians treating cancer patients. Each patient's case is composed of: the patient's disease (i.e. type of cancer), the genetic variants of the disease (i.e. which genes), and some basic demographic information of the patient (i.e. age, gender). Given the condition of a patient, the track proposes two challenges: 1) retrieve the relevant scientific literature about treatments for the specific condition, 2) find relevant clinical trials for which the patient is eligible.

In our participation to the TREC 2019 PM Track, we focused on both tasks with particular emphasis on the clinical trials task. The objective of the work is twofold: (i) we want to evaluate how different query reformulations, validated on previous TREC PM collections [2], affect the results and whether the findings

[1] http://www.trec-cds.org/2019.html

obtained in previous years remain valid; (ii) we want to verify if combining different query reformulations based on expansion and reduction techniques prove effective in such a highly specific scenario.

In this work, we present the experiments we carried out using a fully automated system that: i) performs query reformulations, based on medical knowledge bases [8, 10], to increase the probability of finding relevant documents by adding to or removing from the original query those terms that are related to neoplasm and gene information, respectively; ii) filters out inappropriate clinical trials based on demographic data; iii) performs rank fusion based on the combination of query reformulations that have been validated on previous TREC PM collections.

2 Methodology

The approach proposed comprises four steps, plus an additional step used only for retrieving clinical trials. For each query, the steps are: (2.1) indexing, (2.2) query reformulation, (2.3) retrieval, and (2.4) filtering (only for clinical trials). Then, the rankings obtained by multiple queries can be combined using (2.5) rank fusion.

2.1 Indexing Step

We create the following fields to index clinical trials: `<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, we create the following fields: `<docid>` and `<text>`. As for clinical trials, the `<text>` field contains the entire content of each target document.

2.2 Query Reformulation Step

The approach leverages two types of query reformulation techniques: query expansion and query reduction.

Query expansion: We perform a knowledge-based query expansion. We rely on MetaMap [3], a state-of-the-art medical concept extractor, to extract and disambiguate from each query field all the Unified Medical Language System (UMLS)[2] concepts belonging to the following semantic types:[3] `Neoplastic Process` (*neop*), `Gene or Genome` (*gngm*), and `Cell or Molecular Dysfunction` (*comd*). The

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

gngm and *comd* semantic types are related to the query `<gene>` field, while *neop* is related to the `<disease>` field.

For each disambiguated concept, we consider all its name variants contained into the following knowledge sources: National Cancer Institute Thesaurus[4] (NCIt), Medical Subject Headings[5] (MeSH), SNOMED CT[6] (SNOMEDCT) and UMLS Metathesaurus[7] (MTH). All knowledge sources are manually curated and up-to-date.

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 [6] 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. For instance, consider the TREC 2019 PM topic 1 where the `<gene>` field mentions "BRAF (E586K)". With the reduction process, the `<gene>` field becomes "BRAF". The reduction process mitigates the over-specificity of topics, since the information contained in topics might be too specific compared to that within target documents [11].

2.3 Retrieval Step

We rely on BM25 to retrieve documents. Query terms obtained through query expansion are weighted lower than 1.0 to avoid introducing too much noise in the retrieval process [7].

2.4 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. demographic data (age and gender) do not satisfy the three eligibility criteria aforementioned. 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.

[4] `https://ncithesaurus.nci.nih.gov/ncitbrowser/`
[5] `https://www.ncbi.nlm.nih.gov/mesh/`
[6] `http://www.snomed.org/`
[7] `https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/`

2.5 Rank Fusion Step

We perform rank fusion over the runs obtained with the three most effective query reformulations for clinical trials, whereas we combine the top three query reformulations and the baseline for scientific literature. We adopt the CombSUM [5] technique to perform rank fusion and we normalize scores using min-max normalization.

3 Experiments

For our experiments, we used two different search engine libraries in order to index, retrieve and filter documents: Whoosh[8] for clinical trials, ElasticSearch[9] for scientific abstracts. For the implementation of the BM25 model, we kept the default values $k_1 = 1.2$ and $b = 0.75$ – as we found them to be a good combination for this kind of tasks [1]. For query expansion, we rely on MetaMap to extract and disambiguate concepts from UMLS.

We submitted five runs for each task of the 2019 PM track:

- clinical trials
  1. BM25_baseline refers to BM25 over original queries with no expansion (baseline);
  2. BM25_neop_01_reduc refers to *neop* expansion with expansion weight 0.1 over reduced queries;
  3. BM25_solid_01_orig refers to *solid* expansion with expansion weight 0.1 over original queries;
  4. BM25_solid_01_reduc refers to *solid* expansion with expansion weight 0.1 over reduced queries;
  5. top3_qref_combined refers to the combination of the aforementioned query reformulations using CombSUM.
- scientific literature
  1. BM25_baseline refers to BM25 over original queries with no expansion (baseline);
  2. BM25_neop_01_orig refers to *neop* expansion with expansion weight 0.1 over original queries;
  3. BM25_neop_comd_01_orig refers to the *neop* and *comd* expansions with expansion weight 0.1 over original queries;
  4. BM25_neop_gngm_01_orig refers to the *neop* and *gngm* expansions with expansion weight 0.1 over original queries;
  5. top4_qref_combined refers to the combination of the aforementioned query reformulations plus the baseline using CombSUM.

We summarize the procedure used for each experiment below.

[8] `https://whoosh.readthedocs.io/en/latest/intro.html`
[9] `https://www.elastic.co`

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>`;
- Obtain from extracted 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.

Retrieval

- Adopt the three most effective reformulation strategies from [2];
- Weigh expanded terms with $k = 0.1$;
- Perform a search using expanded queries with BM25.

Filtering

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

Rank fusion

- Perform rank fusion using CombSUM and Min Max normalization over the three most effective query reformulation strategies for clinical trials;
- Perform rank fusion using CombSUM and Min Max normalization over the three most effective query reformulation strategies plus the baseline for scientific literature.

3.1 Results

The organizers of the TREC 2019 PM Track provided the summary of the results in terms of best, median, and worst value for each topic for three evaluation measures: inferred NDCG (infNDCG) [13], precision at 10 (P@10), and R-precision (RPrec).

In Table 1 and 2, we report the median values of the three measures averaged across topics, for the clinical trials and the scientific literature task respectively, as well as the averaged results of the five submitted runs.

For each run, we show a barplot that displays, topic by topic, the difference between the performance of the 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.

measure	median	1	2	3	4	5
infNDCG	0.514	0.619	0.576	0.624	0.594	0.620
RPrec	0.348	0.434	0.413	0.439	0.426	0.438
P@10	0.466	0.505	0.524	0.537	0.532	0.534

Table 1: Overall comparison with average median values of the clinical trials task. The number, from 1 to 5, indicates the run described in Section 3 (i.e. run 2 corresponds to BM25_neop_01_reduc).

measure	median	1	2	3	4	5
infNDCG	0.456	0.475	0.465	0.464	0.474	0.467
RPrec	0.281	0.298	0.298	0.296	0.300	0.300
P@10	0.545	0.512	0.515	0.512	0.505	0.505

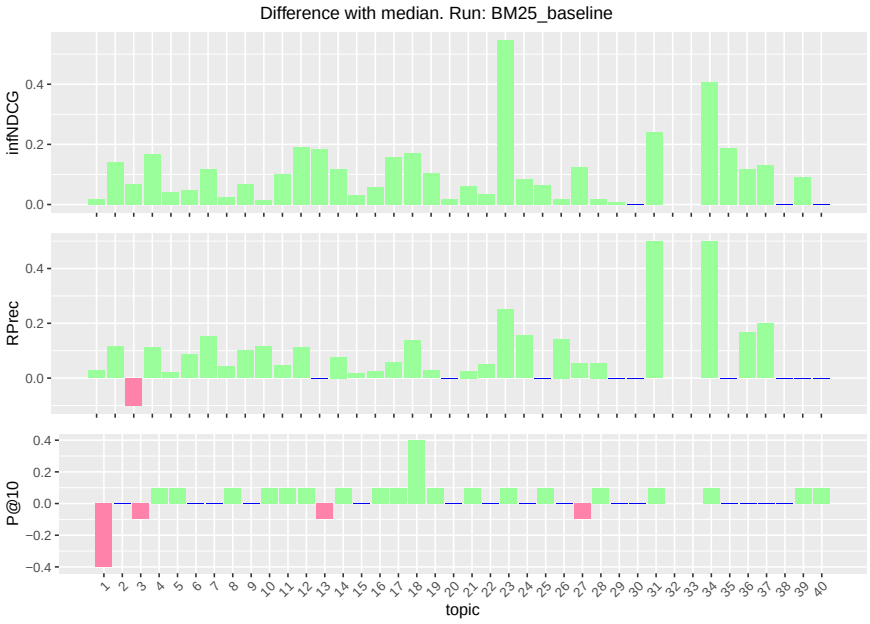
Table 2: Overall comparison with average median values of the scientific literature task. The number, from 1 to 5, indicates the run described in Section 3 (i.e. run 2 corresponds to BM25_neop_01_orig).

The results show that the approach that was studied and evaluated on the test collection of the past two years confirms to be very effective for the clinical trials. For this task, the runs that performed well in previous years confirm a very positive trend and achieve, in most cases, performances that are above the median of the task for all the topics. For scientific abstracts, we see the same moderately good performance in line with previous years. In this task, we could not find any combination of query expansion/reduction with good performances for many topics.

3.2 Comparison with TREC 2019 PM Top Runs

When looking at the detailed analysis of the overview of the TREC PM 2019 task [12], we observe that the performance of our best runs are in the top 10 performing runs per task for all but one performance measure. In particular, for the clinical trials task, the run *solid_01_orig* is the second best run for the infNDCG and the R-prec measure, and the third best run for the P@10. We can confirm also in the case of the scientific literature task that the particular combination of query reformulation and re-weighting that performed well in TREC 2018 is also one of the top performing runs in TREC 2019. For this second task, we observed that our approaches are less effective as other methods in terms of precision in the top elements of the ranking list.

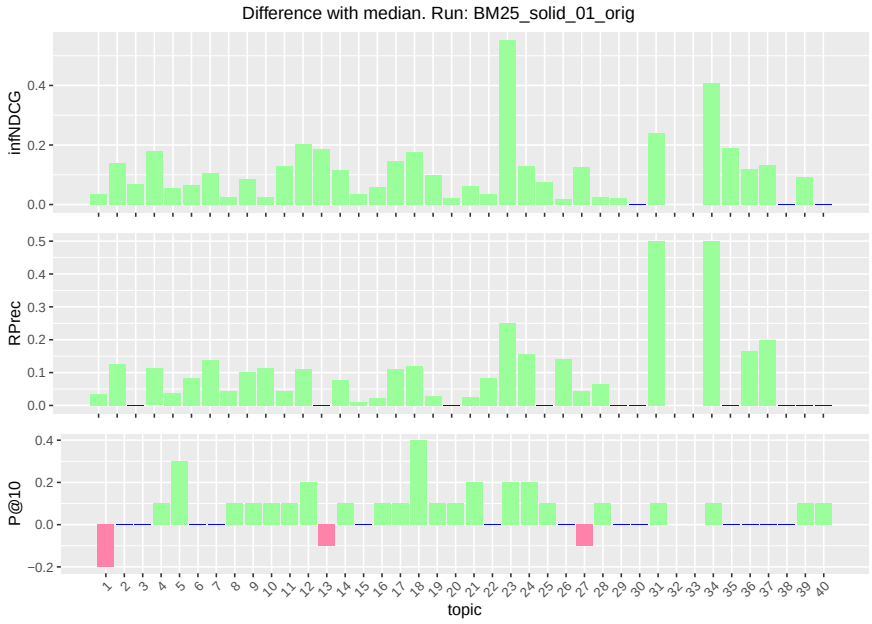
In general, these results are very promising for at least two reasons: the first one is that this particular combination of query expansion and re-weighting shows to be very consistent in terms of effectiveness through the years; the second one is the fact that our approach could be extended to include more sophisticated, and precision-oriented, query boosting approaches [4] or re-ranking techniques [9], thus improving the state-of-the-art.



(a)

(b)

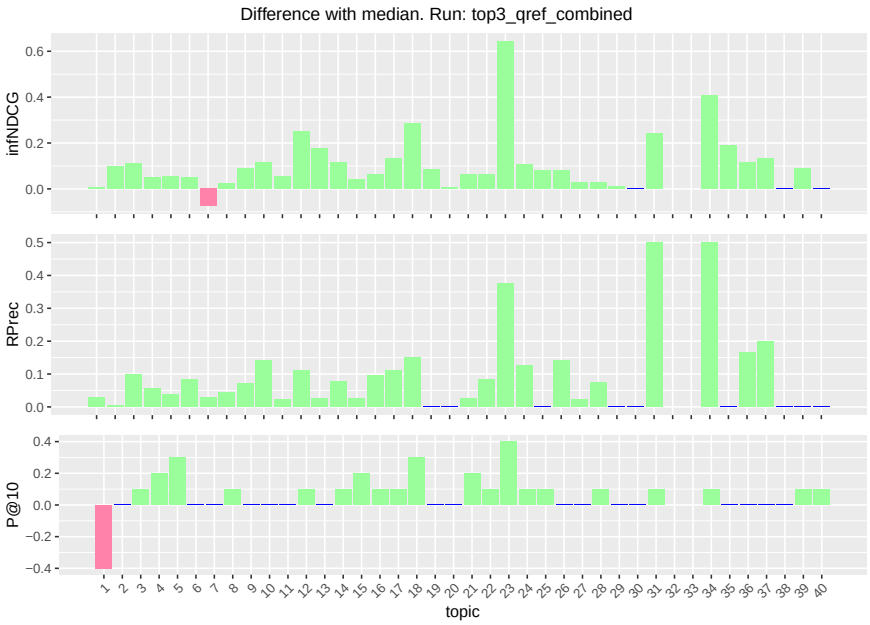
Fig. 1: Topic by topic difference between runs and median values of the clinical trials task.



(a)

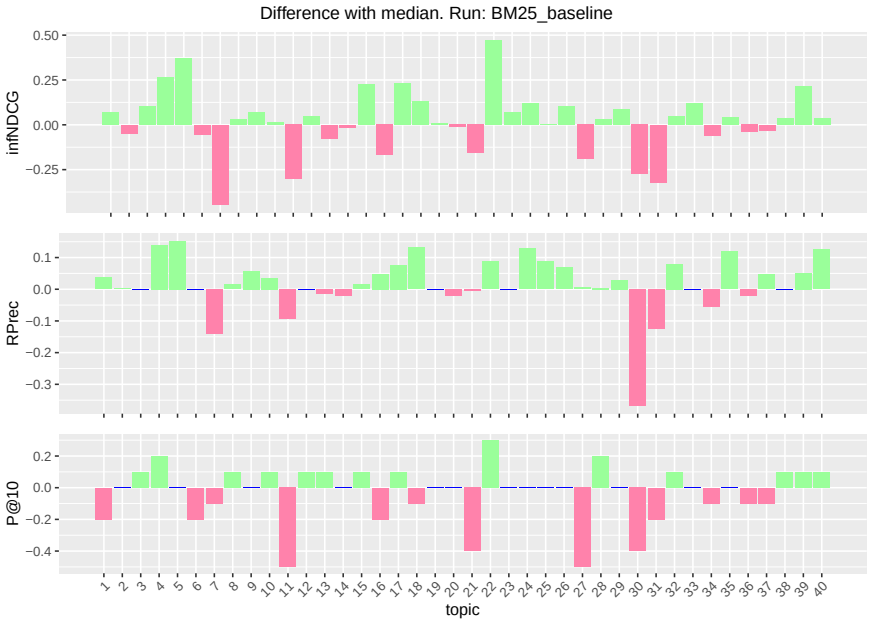
(b)

Fig. 2: Topic by topic difference between runs and median values of the clinical trials task.

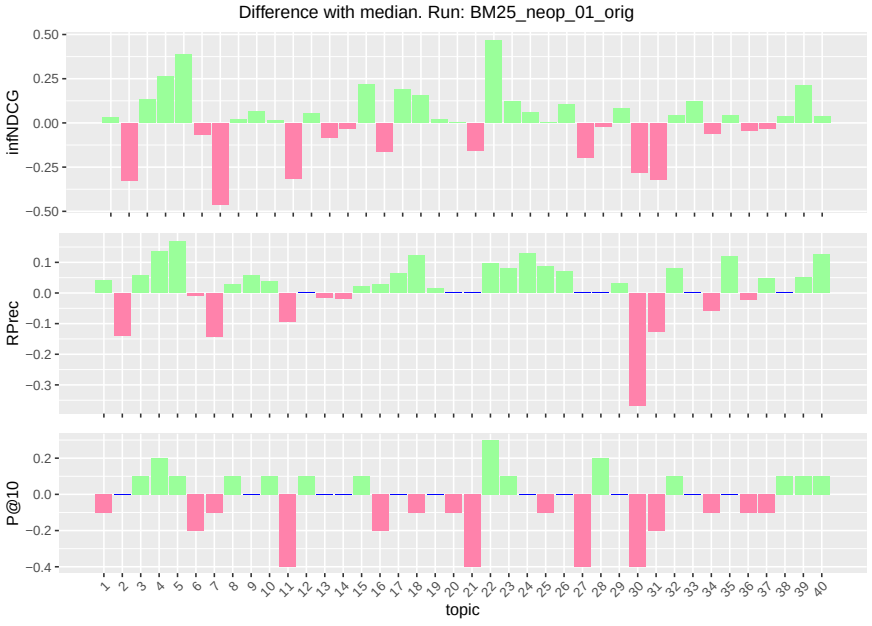


(a)

Fig. 3: Topic by topic difference between runs and median values of the clinical trials task.

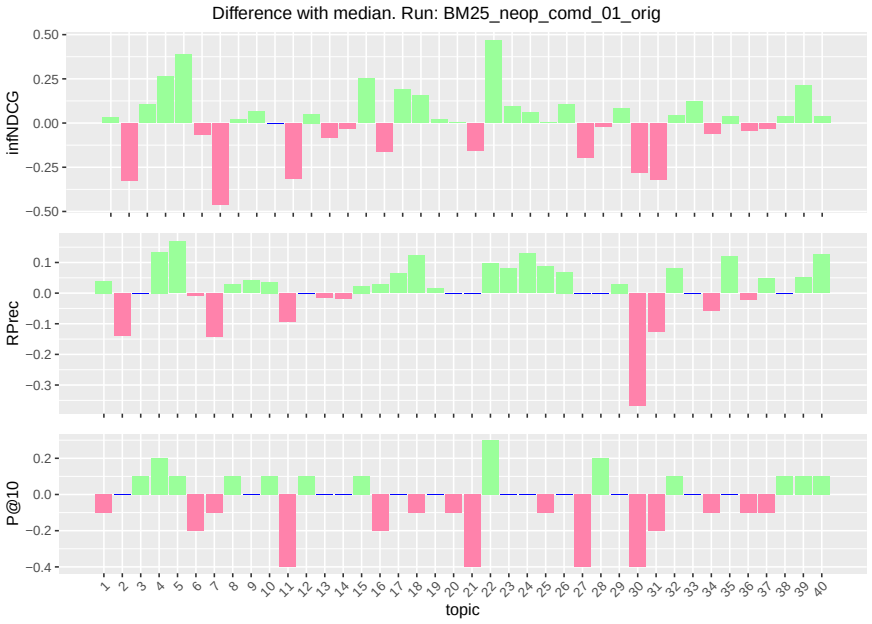


(a)

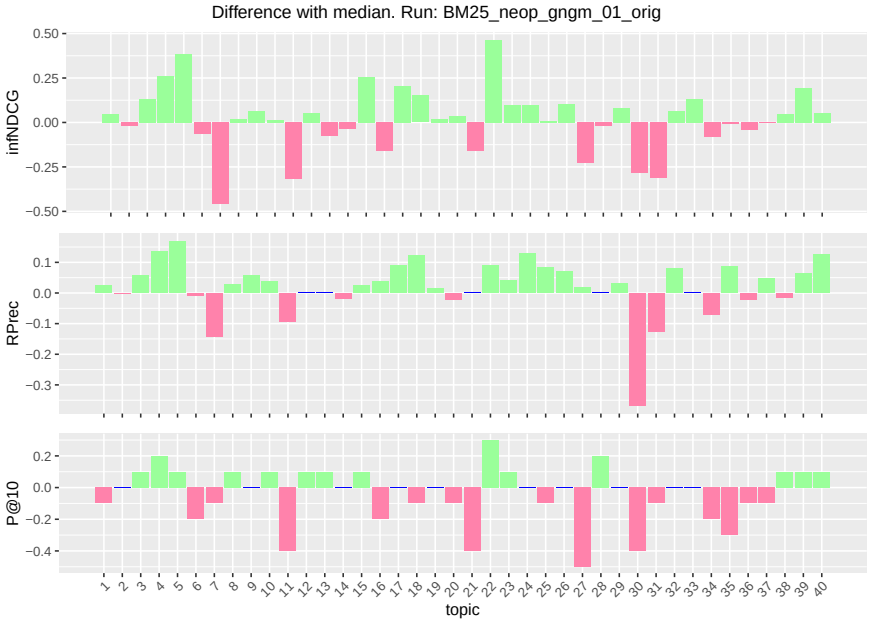


(b)

Fig. 4: Topic by topic difference between runs and median values of the scientific abstracts task.

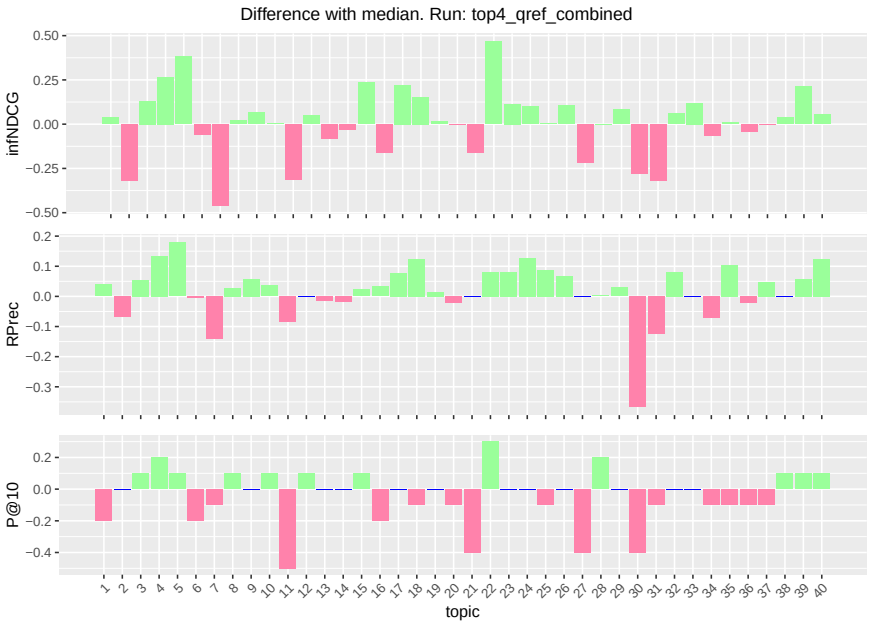


(a)



(b)

Fig. 5: Topic by topic difference between runs and median values of the scientific abstracts task.



(a)

Fig. 6: Topic by topic difference between runs and median values of the scientific abstracts task.

Finally, we stress the fact that the variation of the performance across topics is smaller than any other best performing system (see for example the boxplots in [12]). This is very important since it gives us the opportunity to build a robust baseline independently from the particular set of topics.

4 Final Remarks

In this paper, we presented the results of our second participation in the TREC PM Track. Our objective was the study of knowledge-based query reformulation techniques combined with a rank fusion approach. We relied on the findings from [1, 2] to build our system. For each task, we submitted the top three query reformulations proposed in [2] and we evaluated whether their effectiveness still hold in this new scenario. Furthermore, we investigated on combining different query reformulations to improve the results and build systems more robust to the problem of topic drift.

The analysis of the results confirmed the effectiveness of the query reformulations proposed in [2], especially for the clinical trials task. Additionally, the results obtained by combining top query reformulations with a rank fusion approach highlight the benefits of combining knowledge-based query reformulations in highly specific domains. The combined run provided better results than all the other proposed ones.

As future work, an in-depth investigation on query reformulation techniques could be performed to improve the precision of the system while keeping stable the recall.

Acknowledgements

The work was partially supported by the ExaMode project, as part of the European Union H2020 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 Precision Medicine Track. In: Proceedings of The Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018 (2018)
2. Agosti, M., Di Nunzio, G.M., Marchesin, S.: 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. pp. 973–976. ACM (2019)
3. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium. p. 17. American Medical Informatics Association (2001)

4. Faessler, E., Oleynik, M.: JULIE lab at TREC 2019 precision medicine track. In: Notebook Papers of the Twenty-Seventh Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019 (2019), https://trec.nist.gov/act_part/conference/papers/julie-mug.PM.pdf
5. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. NIST special publication SP **243** (1994)
6. Goodwin, T.R., Skinner, M.A., Harabagiu, S.M.: UTD HLTRI at TREC 2017: Precision medicine track. In: Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017 (2017)
7. Gurulingappa, H., Toldo, L., Schepers, C., Bauer, A., Megaro, G.: Semi-supervised information retrieval system for clinical decision support. In: Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016 (2016)
8. Jimmy, Zuccon, G., Koopman, B.: QUT ielab at CLEF 2018 consumer health search task: Knowledge base retrieval for consumer health search. In: Cappellato, L., Ferro, N., Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2125/paper_203.pdf
9. Liu, X., Li, L., Yang, Z., Dong, S.: SCUT-CCNL lab at TREC 2019 precision medicine track. In: Notebook Papers of the Twenty-Seventh Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019 (2019), https://trec.nist.gov/act_part/conference/papers/julie-mug.PM.pdf
10. Mahmood, A.S.M.A., Li, G., Rao, S., McGarvey, P.B., Wu, C.H., Madhavan, S., Vijay-Shanker, K.: UD_GU_BioTM at TREC 2017: Precision Medicine Track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017. vol. Special Publication 500-324. National Institute of Standards and Technology (NIST) (2017), https://trec.nist.gov/pubs/trec26/papers/UD_GU_BioTM-PM.pdf
11. Oleynik, M., Faessler, E., Sasso, A.M., Kappattanavar, A., Bergner, B., da Cruz, H.F., Sachs, J., Datta, S., Böttinger, E.: HPI-DHC at TREC 2018: Precision Medicine Track. In: Proceedings of The Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018 (2018)
12. Roberts, K., Demner-Fushman, D., Voorhees, E.M., Hersh, W.R., Lazar, A.J., Pant, S., Meric-Bernstam, F.: Overview of the TREC 2019 Precision Medicine Track. In: Proceedings of the Twenty-Eight Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019 (2019), https://trec.nist.gov/
13. Yilmaz, E., Kanoulas, E., Aslam, J.A.: A Simple and Efficient Sampling Method for Estimating AP and NDCG. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 603–610. SIGIR '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1390334.1390437, http://doi.acm.org/10.1145/1390334.1390437