HPI-DHC at TREC 2018 Precision Medicine Track

Michel Oleynik*§, Erik Faessler†§, Ariane Morassi Sasso‡§,
Arpita Kappattanavar‡, Benjamin Bergner‡, Harry Freitas da Cruz‡,
Jan-Philipp Sachs‡, Suparno Datta‡, and Erwin Böttinger‡
*Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria
michel.oleynik@stud.medunigraz.at
†Jena University Language and Information Engineering (JULIE) Lab, Jena, Germany
erik.faessler@uni-jena.de
‡Digital Health Center, Hasso Plattner Institute, Potsdam, Germany
ariane.morassi-sasso@hpi.de
§The first three authors contributed equally to this work.

***Abstract*—The TREC-PM challenge aims for advances in the field of information retrieval applied to precision medicine. Here we describe our experimental setup and the achieved results in its 2018 edition. We explored the use of unsupervised topic models, supervised document classification, and rule-based query-time search term boosting and expansion. We participated in the biomedical articles and clinical trials subtasks and were among the three highest-scoring teams. Our results showed that query expansion associated with hand-crafted rules contribute to better values of information retrieval metrics. However, the use of a precision medicine classifier did not show the expected improvement for the biomedical abstracts subtask. In the future, we plan to add different terminologies to replace hand-crafted rules and experiment with negation detection.**

I. Introduction

According to the U.S. National Research Council, precision medicine aims to provide the best possible care for a patient by tailoring the treatment to its individual characteristics [1]. Efficient access to the existing scientific and medical literature is therefore a critical instrument to obtain treatment information related to the patient's profile. To progress in this field, the National Institute of Standards and Technology (NIST) has organized the TREC Precision Medicine (TREC-PM) track since 2017.

This paper describes the participation of the "hpi-dhc" group. The team was formed by members of three different institutions with different backgrounds, including computer science and medicine. TREC-PM was held in 2018 with the same goal as the previous edition: given structured data related to an oncology patient case (henceforth denoted "topic"), retrieve relevant (1) Biomedical Abstracts (BA) from PubMed, and (2) Clinical Trials (CT) from ClinicalTrials.gov.

The rationale behind the usage of two datasets is a cascaded search strategy. First, the existing literature is scanned for a disease and its connection to a specific genetic profile. If this search does not bring up the desired information, the patient can then potentially be enrolled on a clinical trial. The ideal precision medicine search engine would thus propose relevant literature articles first and then match clinical trials in case the former did not yield relevant results.

```
<topic number="38">
  <disease>cholangiocarcinoma</disease>
  <gene>IDH1</gene>
  <demographic>50-year-old male</demographic>
</topic>
```

Fig. 1. Example of a 2018 TREC-PM topic.

Each topic to be queried had a *disease*, *gene* and *demographic* component (Figure 1), which gave information about the type of cancer, biomarker, age and sex of the patient, respectively. The total amount of topics increased from 30 in 2017 to 50 in 2018 and, in the latter, half of them were related to melanoma. More information about the challenge, the full content of the topics, and the guidelines are available online at https://trec-cds.appspot.com/2018.html.

This paper proceeds as follows. Section II provides detailed information on the methods we applied, including a description of the experimental framework, query expansion and boosting strategies, the usage of reference standards, and rules hand-crafted for the tasks. Section III then describes specific strategies used for processing biomedical abstracts, including an Unstructured Information Management Architecture (UIMA) pipeline, the usage of Topic Modelling (TM) and Term Frequency - Inverse Document Frequency (TF-IDF) analysis to discover relevant boosting keywords, and their consolidation into a Precision Medicine (PM) classifier. Subsequently, Section IV outlines the results obtained when applying the aforementioned strategies. Finally, in Section V we discuss directions for further areas of development and conclude in Section VI with a review of these results in the context of information retrieval for precision medicine.

II. Methods Overview

A. Experimental Framework

We built our work on top of an existing Java framework proposed in 2017 by the Medical University of Graz [2]. With its aid, we indexed TREC-PM data on an Elasticsearch (ES)[1] 5.4.0 instance and performed query-time experimenta-

[1]https://www.elastic.co

```
"bool": {
    "must": [
        {{biomedical_articles/disease.json}},
        {{biomedical_articles/gene.json}}
    ],
    "should": [
        {{biomedical_articles/extra.json}},
        {{biomedical_articles/chemotherapy.json}},
        {{biomedical_articles/cancer.json}},
        {{biomedical_articles/dna.json}},
        {{biomedical_articles/positive_boosters.json}},
        {{biomedical_articles/negative_boosters.json}},
        {{biomedical_articles/pm.json}}
    ],
    "must_not": [
        {{biomedical_articles/non_melanoma.json}}
    ]
}
```

Fig. 2. Example of a template with sub-templates.

tion leveraging the ES query language. We also implemented an UIMA[2] pipeline to preprocess biomedical abstracts, described in Section III-A. We released the code for this year's experiments at https://github.com/hpi-dhc/trec-pm. Internally produced data, submitted runs, and additional graphs are available on figshare[3].

For the experiments, our first step was to add support for dynamic query decorators in the framework, which allowed on-the-fly construction of disjunction max queries (`dis_max`) containing synonyms and hypernyms with different weights. This enabled us to improve recall without causing a corresponding loss in precision, which was reported as a major limitation by the framework creators.

On the technical side, we also added support for sub-templates, which improved modularity and code reuse. Templates and sub-templates were created as JSON files (see Figure 2) and helped us to compose different experiments more easily.

B. Query Expansion

We used the framework to properly expand the *disease* and *gene* fields from the TREC-PM topics and thus improve recall. We enriched the *disease* with its preferred term, synonyms, and hypernyms provided by Lexigram[4], a proprietary API based on the Systematic Nomenclature of Medicine - Clinical Terms (SNOMED CT), the Medical Subject Headings (MeSH) and the International Classification of Diseases (ICD). This allowed us to match e.g. documents mentioning "bile duct carcinoma" to the topic "cholangiocarcinoma" (see Table I). We further enriched the *gene* with its description and synonyms provided by the National Center for Biotechnology Information (NCBI) gene list[5]. For example, documents that mentioned the gene "NS7" would be correctly matched to Topic 1, whose gene is "BRAF" (see Table II).

For both the *disease* and *gene* dimensions, we gave the highest weight to the original topic term and lower weights to the preferred term, synonyms, and hypernyms (see Table III).

[2]https://uima.apache.org

[3]https://figshare.com/projects/TREC_PM_2018_Data_hpi-dhc_/56882

[4]https://www.lexigram.io

[5]ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

TABLE I
EXAMPLE OF DISEASE EXPANSION PROVIDED BY THE LEXIGRAM API

Original term	cholangiocarcinoma
Preferred term	carcinoma of cervix
Synonyms	cholangiocellular carcinoma
	bile duct carcinoma
	bile duct adenocarcinoma
Hypernyms	malignant neoplasm of digestive system
	abdominal mass
	epithelial neoplasm
	disorder of biliary tract
	neoplasm of digestive organ
	finding of biliary tract
	gastrointestinal tract finding

TABLE II
EXAMPLE OF GENE EXPANSION PROVIDED BY THE NCBI GENE LIST

Original term	BRAF
Description	B-Raf proto-oncogene, serine/threonine kinase
Synonyms	B-RAF1
	B-raf
	BRAF1
	NS7
	RAFB1

TABLE III
WEIGHT VALUES FOR DISEASE AND GENE IN BOTH SUBTASKS

Expansion type		**Biomedical articles**	**Clinical trials**
Disease	Original	1.0	1.0
	Preferred	0.1	0.1
	Synonyms	0.1	0.1
	Hypernyms	-	0.6
Gene	Original	1.0	1.0
	Description	0.1	0.1
	Synonyms	0.7	0.1

As a last resource, we added remaining documents with a `match_all` clause with negative boost. This allowed partial matches only on disease or gene and ensured that every topic had at least 1,000 results, the limit imposed by the challenge organizers.

C. Query Boosting

We also leveraged the framework to boost documents related to the task subdomain. Following previously tested strategies, we prioritize documents mentioning e.g. the following positive keywords:

- Oncology: "cancer", "carcinoma", "tumor"
- Precision medicine: "treatment", "prevention", "prognosis", "survival", "outcome", "resistance"
- Genetics: "gene", "genotype", "DNA", "base"
- Chemotherapy suffixes: "*mab", "*nib", "*cin", "*one", "*ate", "*mus", "*lin"

For the BA subtask, we further boosted papers from the ASCO and AACR conferences on oncology, as this was deemed more relevant by the organizers due to their publications being more recent. Conversely, we downgraded documents related to *in vitro* research, viz. containing negative keywords such as "tissue" and "cell".

D. Reference Standard

For the 2018 TREC-PM challenge, the organizers provided access to the official 2017 Gold Standard (GS)[6], which has relevance assessments for 22,642 and 13,441 query-document pairs from the BA and CT tasks, respectively [3]. It also contained the annotations that led to the final relevance assessments, including whether a given document was considered "Animal PM", "Human PM", or "Not PM". We thus leveraged it to (1) test hypotheses; (2) debug results; (3) find optimal weight values; (4) model topics and analyze TF-IDF data from PM biomedical articles (see Section III-B); and (5) build a PM classifier for biomedical articles (see Section III-C).

As topics differed from the previous edition, we also created an internal gold standard for 2018 containing 336 and 141 query-document pairs for the biomedical articles and clinical trials tasks, respectively.

E. Hand-crafted Rules

Based on the manual assessment of results in the GS, we enriched queries with additional hand-crafted rules for non-melanomas, solid tumors, and gene families.

Firstly, we noticed that a high number of results for topics about "melanoma" would match biomedical articles and clinical trials about "non-melanoma". We overcame this situation by manually adding an exact query clause to exclude such results. This was aimed at improving precision on the relevant topics.

Secondly, we realized that a large part of relevant clinical trials in the 2017 reference standard did not mention the exact topic disease, but would rather prefer an umbrella term, such as "solid tumor". Since such concepts were not included in the terminologies we used, we built a simple query decorator that would add "solid" as a hypernym if the *disease* did not mention "lymphoma" or "leukemia". This was geared towards improving recall.

Thirdly, we observed that clinical trials would commonly not mention the exact gene from the topic, but its family (e.g. "BRCA" instead of "BRCA2"). Instead of integrating terminology resources with such information, we explored a straightforward and naive approach based on regular expressions to generate proper term variations and thus improve recall. Using the pattern `([0-9]{1,2}[A-Z]{0,2}|R[0-9]{0,1})$`, we removed up to two trailing digits (followed or not by up to two letters) or a trailing `R` (denoting "receptor") optionally followed by a digit. This accounts e.g. for the following substitutions:

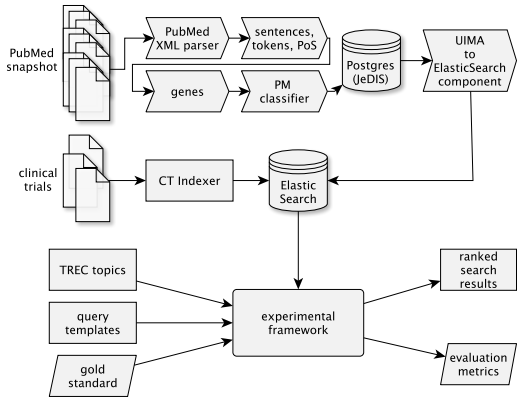

Fig. 3. An overview of our complete experimental setup.

- BRCA2 → BRCA
- TP53 → TP
- CDK6 → CDK
- FGFR1 → FGF
- EGFR → EGF
- PIK3CA → PIK

III. Additional Methods for Biomedical Abstracts

Given the large amount of documents in the biomedical abstracts subtask and its unstructured nature, we adapted existing components to process such data in an efficient way and explored new tools to better understand the idea of "precision medicine" captured by annotators in the gold standard. We therefore describe such efforts in this dedicated section.

A. UIMA Pipeline

The main building block of the experimental architecture for BA was a linguistic UIMA preprocessing pipeline connected to the ES instance and the experimental framework as shown in Figure 3. The right-arrow elements stand for UIMA components, sometimes aggregating multiple primitive components. The rectangle with rounded edges represents the experimental framework (described in Section II-A) which is the heart of our research effort. The other rectangles connected to the experimental framework depict the input and output data that are always required or delivered. The parallelograms represent the case where evaluation data is available to indicate that the experimental framework can run searches on ES as well as evaluate the results.

We used the UIMA-based JCoRe repository [4] to read the 2017 PubMed snapshot delivered by the challenge organizers, to recognize gene and organism mentions, apply our precision medicine document classifier (see Section III-C) and, finally, to index the output of an Natural Language Processing (NLP) pipeline results into ES. For the recognition of gene mentions we employed BANNER[7] which is available as a JCoRe

[6]https://trec.nist.gov/data/precmed2017.html

[7]http://banner.sourceforge.net

component[8]. The BANNER gene model was trained on the complete BioCreative II GM[9] train and test data.

As depicted in Figure 3, the JCoRe components form a processing sequence called *pipeline* in UIMA: the output of one component may serve as input for subsequent components. Since the gene tagger requires linguistic information such as sentence and token segmentation of the text and parts of speech for the tokens, the basic linguistic processing components run before the gene tagger. Those are actually multiple components but they are shown aggregated in Figure 3 for simplicity reasons. We indexed in ES the document ID, title, abstract, keywords, MeSH headings, gene mentions found by BANNER and the precision medicine label as predicted by the PM classifier (see Section III-C).

The resulting UIMA text annotations were stored in a PostgreSQL database leveraging the JeDIS system [5]. With it, we could store all annotation levels (e.g. tokens, sentences, or genes) separately from each other, which allowed individual updates and eased experimentation. In our case, we could run development versions of the PM classifier and store the results in the JeDIS annotation database without the need to recreate the other annotations.

B. Topic Modelling and TF-IDF

As manual keyword selection for boosting (see Section II-C) is time-consuming and requires curation by experts, we explored automated approaches that could exploit the knowledge already encoded in the GS to identify candidate words for boosting.

We first leveraged the official annotations linked to the 2017 GS (as described in Section II-D) to identify topics. Topic Modelling (TM) is an unsupervised technique to build probability distributions between so-called *topics* and the word types observed in a set of input documents. Topics form an intermediate layer between a document d and its contained word types $w \in V$ where V is the vocabulary of words taken into account. A document covers a set of topics where each topic t defines an *a posteriori* distribution $p(w|t)$ of words occurring in t [6], [7]. We applied Latent Dirichlet Allocation (LDA) [8] using the MALLET[10] 2.0.9 toolkit. TM found topics whose top-words strongly matched our intuition of PM.

In a similar effort, we experimented with TF-IDF analysis in the 2017 GS to obtain discriminant keywords of relevant PM biomedical articles. We thus looked for words that had a high TF-IDF score for PM and a low score for Not-PM, further selected as candidate "positive keywords". Conversely, we looked for terms with a high TF-IDF score for Not-PM and a low TF-IDF score for PM and selected them as candidates for "negative keywords".

In order to craft the final keyword lists, we mixed the distinct candidate terms obtained from both approaches (i.e. LDA and TF-IDF) and created experiments with different combinations of them to analyze which ones improved evaluation metrics the most. We then replaced our positive and negative boosters with the ones providing a real improvement on metrics (see Table IV).

TABLE IV
Positive and negative boosters

Positive	gefitinib treatment survival prognostic clinical prognosis therapy outcome resistance Gleason targets
Negative	pathogenesis tumor cell development model tissue mouse specific staining dna case

C. PM Classifier

We also explored a supervised approach to automatically learn the intuition behind "precision medicine" and improve ranking with a so-called PM classifier.

For this task, we computed word TF-IDF estimates of document tokens from the TREC-PM 2017 gold standard using the SecondString[11] library. The documents' token TF-IDF values were used as features in a bag-of-words approach (A). As precision medicine often revolves around specific genes, we also added automatic gene mention tagging via BANNER to our preprocessing as described in Section III-A. We added the number of found genes in a document as well as the name of the genes as they appeared in the text (B). Additionally, we added the names of organism text mentions as detected by the LINNEAUS tagger [9] (C) and, if available, the major MeSH descriptor names of the document (D) as features. Some documents were inconsistently tagged for PM in the gold standard for different query topics; in such cases we assigned the document the PM gold label.

We then built a Maximum Entropy (a.k.a logistic regression) classifier on top of these four feature groups. Figure 4 illustrates how the features were extracted from a document and placed into the classifier to obtain the PM or the Not-PM label for the input document. A stratified ten-fold cross-validation on the 2017 TREC-PM gold standard yielded a classification accuracy of 75%. The classifier was eventually trained on the complete GS and applied to the whole PubMed snapshot. For each document, its label was stored in the search index and used at query time to boost documents that had been classified to be relevant to precision medicine.

We additionally tested the impact of each feature group on classification accuracy when compared to a baseline set as TF-IDF only (A). We thus designed four experiments[12], three with one feature group disabled and one with all feature groups, all including the TF-IDF baseline.

We performed paired Student's t-tests between each combination of the experiments (resulting in six tests since the direction of the test does not matter). For each comparison, the

[8]https://github.com/JULIELab/jcore-base

[9]http://biocreative.sourceforge.net/biocreative_2_gm.html

[10]http://mallet.cs.umass.edu/

[11]http://secondstring.sourceforge.net

[12](A)+(B)+(C), (A)+(B)+(D), (A)+(C)+(D), and (A)+(B)+(C)+(D).

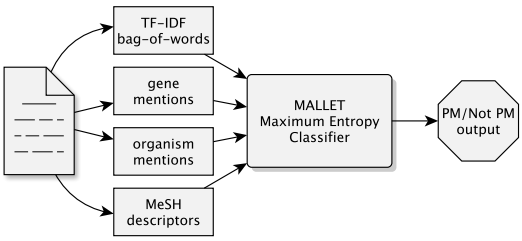

Fig. 4. Features and method used for classifying a document as "precision medicine".

TABLE V
DESCRIPTION OF BIOMEDICAL ARTICLES RUNS

	hpipub				
Strategies	**base**	**common**	**none**	**class**	**boost**
Positive/negative boosters	Y	Y	Y	Y	Y
Disease/gene exact match	Y	Y	Y	Y	Y
Non-melanoma rule	Y	Y	Y	Y	Y
Disease preferred term	N	Y	Y	Y	Y
Disease synonym	N	Y	Y	Y	Y
Gene description	N	N	Y	Y	Y
Gene synonym	N	Y	Y	Y	Y
PM classifier	N	N	N	Y	Y
PM classifier boost	N	N	N	N	Y

test input was the ten accuracy scores obtained during cross-validation for each experiment. After performing the Holm-Bonferroni correction with a significance level of $\alpha = 0.05$, the following comparisons showed significant differences:

- no gene features vs. no MeSH features ($p < 0.00001$)
- no gene features vs. no organism features ($p < 0.0001$)
- no gene features vs. all features ($p < 0.001$)

In all comparisons, the experiment excluding the gene features showed a lower accuracy than the other experiment. The gene features group was the only one that brought an actual improvement from a statistical point of view. The omission of the gene features had the largest impact and caused the ten-fold cross-validation accuracy to drop to 73%. Other runs varied only slightly from the score achieved with all features. Since the other features did not hurt, we left them activated in the hope of better generalization on unseen data.

IV. RESULTS

We submitted only automatic runs using differently weighted combinations of the strategies presented before. Tables V and VI show the approaches we have applied for each of the five runs for the biomedical articles and clinical trials tasks, respectively.

A. Biomedical Abstracts

Table VII shows our overall results for the biomedical abstracts subtask and Figure 5 shows boxplots over all topics comparing the five submitted runs to the average best and median results (over all participants). The `hpipubcommon` run had the best infNDCG (0.5605), P@10 (0.7060), and R-Prec (0.3658), on a tie with the `hpipubnone` run for the infNDCG and P@10 metrics.

Run `hpipubbase` had slightly lower average for all metrics, which might suggest a positive effect of query expansion, as this was the only run not using it. Moreover, we could not show a benefit of using a PM classifier, as there was no large differences between the runs `hpipubnone` (not using the classifier), `hpipubclass` (using the classifier), and `hpipuboost` (boosting the classifier).

Figure 7 in the appendix shows the results for each topic. Considering P@10 for the `hpipubcommon` run, topics 35 and 28 had the largest negative difference (-0.1000) to the median of all participant runs for this topic (not shown in the

TABLE VI
DESCRIPTION OF CLINICAL TRIALS RUNS

	hpict				
Strategies	**base**	**common**	**boost**	**phrase**	**all**
Positive/negative boosters	Y	Y	Y	Y	Y
Disease/gene exact match	Y	Y	Y	Y	Y
Age/sex match	Y	Y	Y	Y	Y
Disease preferred term	N	Y	Y	Y	Y
Disease synonym	N	Y	Y	Y	Y
Disease hypernym	N	Y	Y	Y	Y
Gene description	N	N	N	N	Y
Gene synonym	N	Y	Y	Y	Y
Non-melanoma rule	Y	Y	Y	Y	Y
Solid tumor rule	N	N	Y	Y	Y
Gene family rule	N	N	Y	Y	Y
Phrase matching	N	N	N	Y	N

TABLE VII
BIOMEDICAL ABSTRACTS: OVERALL RESULTS

Run	**infNDCG**	**P@10**	**R-Prec**
hpipubboost	0.5574	0.7040	0.3656
hpipubnone	**0.5605**	**0.7060**	0.3648
hpipubbase	0.5235	0.6920	0.3481
hpipubclass	0.5554	0.6980	0.3547
hpipubcommon	**0.5605**	**0.7060**	**0.3658**

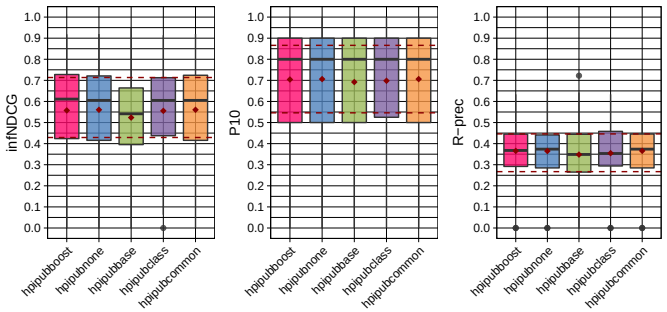

Fig. 5. Biomedical Abstracts: boxplots comparing our runs to the average best and median results.

TABLE VIII
CLINICAL TRIALS: OVERALL RESULTS

Run	infNDCG	P@10	R-Prec
hpictall	**0.5545**	0.5340	0.3964
hpictphrase	0.5484	**0.5400**	**0.4081**
hpictboost	0.5536	0.5340	0.3962
hpictcommon	0.5374	0.5340	0.3953
hpictbase	0.4891	0.4880	0.3715

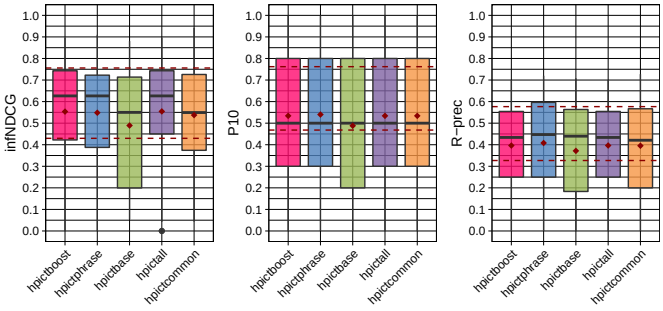

Fig. 6. Clinical Trials: boxplots comparing our runs to the average best and median results.

graph). Conversely, topics 35 (-0.0386) and 11 (-0.0263) had the largest negative differences to the median infNDCG.

B. Clinical Trials

Table VIII shows our overall results for the clinical trials subtask. Figure 6 shows boxplots over all topics comparing the five submitted runs to the average best and median results (over all participants). The `hpictphrase` run had the best P@10 (0.5400) and R-Prec (0.4081), while the `hpictall` run had the best infNDCG (0.5545).

Run `hpictbase` had slightly lower average for all metrics and consistently larger variance, probably due to the absence of query expansion and in tandem with BA results. Moreover, run `hpictcommon` had also a slightly lower median infNDCG. As it was the only run without gene family and solid tumor expansion (as described in Section II-E), this might suggest some benefit from these approaches. Our best runs `hpictphrase` and `hpictall` used almost the same strategies, except for "gene description" that was only used by `hpictall` and "phrase matching" that was exclusive to `hpictphrase`. This might suggest a benefit of using exact matching for CT.

Figure 8 in the appendix shows the results for each topic. Considering P@10 for the `hpictphrase` run, topics 40 and 33 had the largest negative difference (-0.3000) to the median of all participant runs for this topic (not shown in the graph). Conversely, topics 33 (-0.1133) and 11 (-0.0993) in the `hpictall` run had the largest negative differences to the median infNDCG.

V. DISCUSSION

Our work has some important limitations.

Firstly, despite the importance of query expansion and weighting, relative weights were found with a local greedy search only. A more comprehensive approach would be to perform grid search on all possible combinations. However, due to the high-dimensional problem, a naive approach is not feasible and thus some guided search (e.g. employing genetic algorithms) may be necessary.

Secondly, the terminological resources we employed were not fully comprehensive and therefore would miss some important parent concepts, as well as gene families. We tried to overcome that with hand-crafted rules, but a more robust approach would be to integrate other terminologies containing the necessary mappings. For example, we noticed that the NCI Thesaurus includes a predicate `Neoplasm_Has_Special_Category` for solid tumors.

Thirdly, some clinical trials included negated assessments in the inclusion criteria (e.g. "no prior history of breast cancer") instead of explicitly expressing them as exclusion criteria. Moreover, topics 21 and 22 required discriminating between "no" and "extensive" "tumor infiltrating lymphocytes". In order to proper tackle such cases, a proper mechanism for negation detection should be employed, which was not explored by our team. We did, however, manually addressed the prevalent case of "non-melanoma", as described in Section II-E.

Lastly, we trained the PM classifier only on biomedical abstracts and did not evaluate its overlap with the semi-automated keyword selection approach. Also, we used only the binary classifier output for filtering, while we could have used the probability value itself for improved ranking. We believe addressing such issues could lead to a fully-automated boosting mechanism that only depends on training data, thus turning TM and TF-IDF analysis unnecessary.

VI. CONCLUSION

Our work explored weighted query expansion with terminological resources for synonyms and hypernyms, keyword-based query boosting with terms obtained semi-automatically from topic modelling and TF-IDF analysis, a "precision medicine" classifier, and hand-crafted rules for issues not easily solved otherwise.

Weighted query expansion showed that it is possible to improve recall without a loss in precision and thus provided the most positive impact in our experiments. Associated with rules to infer the gene family and detect solid tumors, it further impoved clinical trials metrics. Furthermore, we showed that exact matching improves P@10 and R-Prec in the clinical trials subtask.

Results using a supervised PM classifier proved inconclusive as there was an overlap with the manual keyword boosting strategy. Nevertheless, we proved that a gene tagger does improve the accuracy of such classifier.

Overall, we had the top-performing P@10 and second best infNDCG and R-Prec in the biomedical articles subtask. Considering clinical trials, our group had the top-performing infNDCG. Therefore, next steps would involve building upon the existent successful strategies.

ACKNOWLEDGMENT

Our work is funded by the Brazilian National Research Council - CNPq (project number 206892/2014-4); the German Bundesministerium für Bildung und Forschung (BMBF) under grant no. 01ZZ1803G as well as by the Deutsche Forschungsgemeinschaft (DFG) as part of the CRC 1076 AquaDiva; and by the Hasso Plattner Institute (HPI).

REFERENCES

[1] U. N. R. Council, *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: The National Academies Press, 2011.

[2] P. López-García, M. Oleynik, Z. Kasáč, and S. Schulz, "TREC 2017 Precision Medicine - Medical University of Graz," *TREC, Gaithersburg, MD*, 2017.

[3] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bredrik, A. J. Lazar, and S. Pant, "Overview of the TREC 2017 Precision Medicine Track," in *TREC 2017 — Proceedings of the 26th Text REtrieval Conference. Gaithersburg, Maryland, USA, November 15-17, 2017*, ser. NIST Special Publication, E. M. Voorhees and A. Ellis, Eds., no. SP 500-324. Gaithersburg/MD: National Institute of Standards and Technology (NIST), 2017, pp. 1–13. [Online]. Available: http://trec.nist.gov/pubs/trec26/papers/Overview-PM.pdf

[4] U. Hahn, F. Matthies, E. Faessler, and J. Hellrich, "UIMA-based JCoRe 2.0 goes GitHub and Maven Central: State-of-the-art software resource engineering and distribution of NLP pipelines," in *LREC 2016 —Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*. Paris: European Language Resources Association (ELRA-ELDA), 2016, pp. 2502–2509.

[5] E. Faessler and U. Hahn, "Annotation level document management with JeDIS," in *DocEng 2018 —Proceedings of the 18th ACM Symposium on Document Engineering. Halifax, Nova Scotia, Canada, August 28-31, 2018*. New York/NY: Association for Computing Machinery (ACM), 2018.

[6] J. Boyd-Graber, Y. Hu, and D. Minmo, *Applications of Topic Models*, ser. Foundations and Trends(r) in Information Retrieval Series. Now Publishers, 2017.

[7] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[9] M. Gerner, G. Nenadi?, and C. M. Bergman, "Linnaeus: A species name identification system for biomedical literature," *BMC Bioinformatics*, vol. 11, p. #85, 2010.

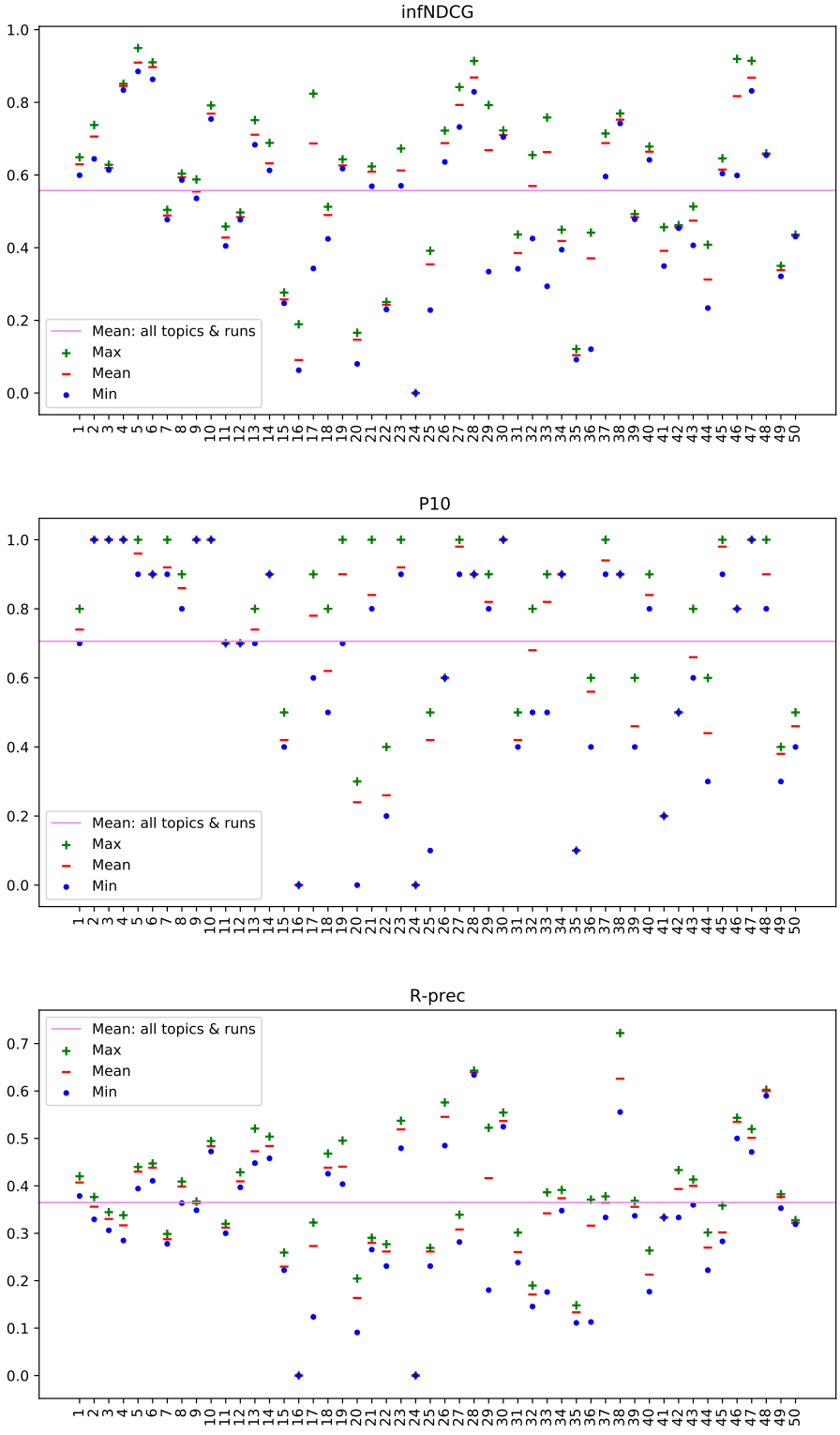

Fig. 7. Biomedical Abstracts: maximum, mean, and minimum infNDCG, P@10, and R-Prec per topic of the five submitted runs, compared to the average median over all participants runs.

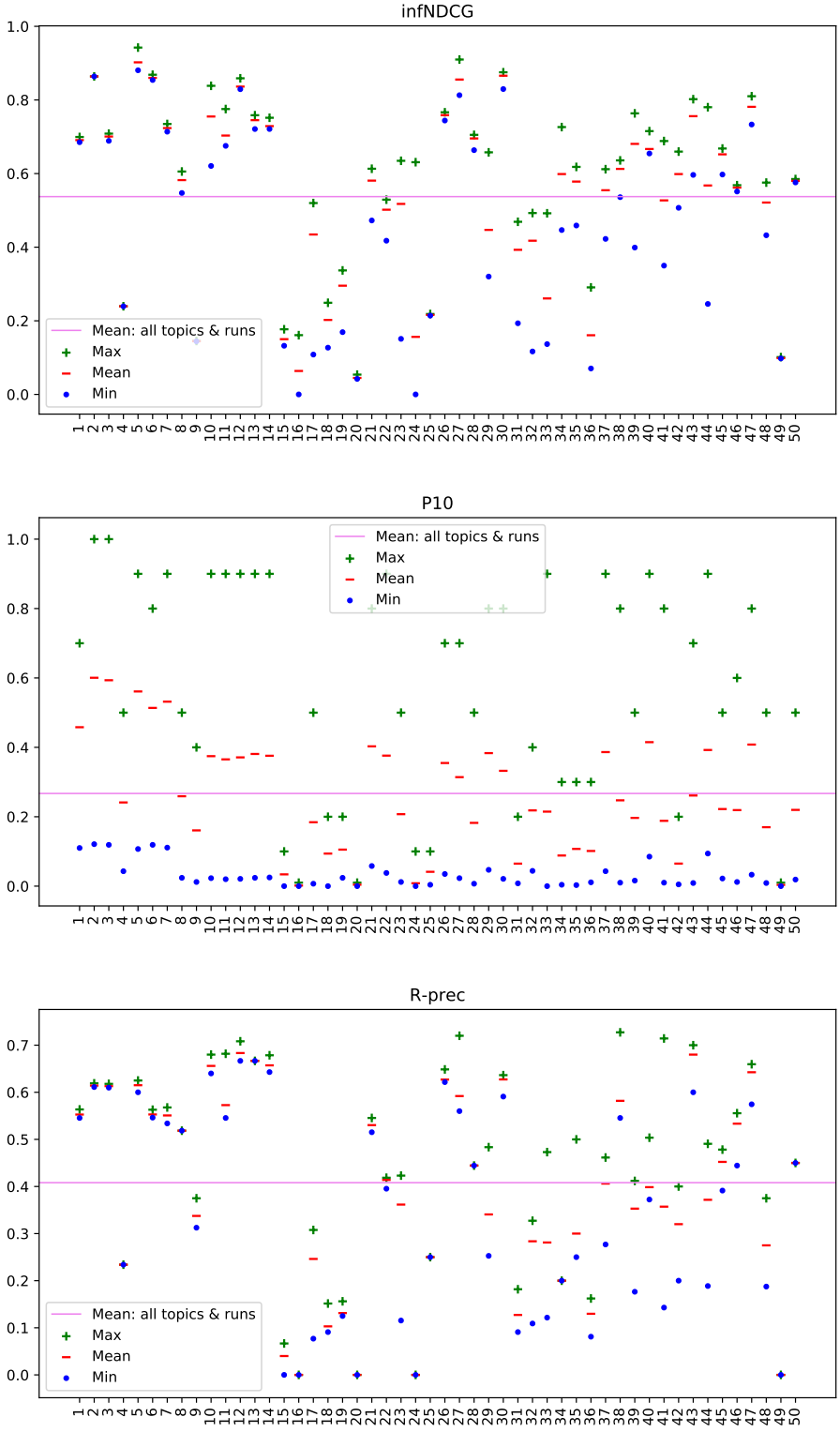

Fig. 8. Clinical Trials: maximum, mean, and minimum infNDCG, P@10, and R-Prec per topic of the five submitted runs, compared to the average median over all participants runs.