Designing retrieval models to contrast precision-driven ad hoc search vs. recall-driven treatment extraction in Precision Medicine

Déborah Caucheteur[a,b], Emilie Pasche[a,b], Julien Gobeill[a,b], Anaïs Mottaz[a,b], Luc Mottin[a,b,c], Patrick Ruch[a,b]

[a] HES-SO / HEG Geneva, Information Sciences, Geneva, Switzerland

[b] SIB Text Mining, Swiss Institute of Bioinformatics, Geneva, Switzerland

[c] Department of Microbiology and Molecular Medicine, Faculty of Medicine, University of Geneva, Geneva, Switzerland

contact: {deborah.caucheteur;emilie.pasche}@hesge.ch

Abstract

The TREC 2019 Precision Medicine Track repeats the general structure and evaluation of the 2018 track. Our team participated in both tasks of the track, relative to scientific abstracts and clinical trials. 40 topics where patient data are given (demographic data, disease, gene and genetic variant) were available for this competition. The aim was to retrieve scientific abstracts and clinical trials of interest regarding a topic, modelling the description of a clinical case. In the first task, we aim at retrieving scientific abstracts introducing some relevant treatments for a given case. Our system is first based on the collection of a large set of abstracts related to a particular case using various strategies such as search with keywords within abstracts, search with normalized entities within annotated abstracts and the linear combination of various queries. We then apply different strategies to re-rank the resulting scientific abstracts set. In particular, we tested two strategies to re-rank the abstracts set in order to have a large variety of treatments returned in the top articles. Almost two thirds of the top-10 returned documents are judged relevant, while nearly a quarter of the relevant treatments is returned in the top-10 abstracts. The second task aims at retrieving some clinical trials for which patients are eligible. Criteria used to determine the eligibility of patients are those found in the topics. Information such as trial location or status of clinical trials, which are important from a patient's point of view, are questionably not used in these topics. Several strategies have been tested, relaxing of constraints (data required or not), expansion of information requests thanks to synonyms or regex, and retrieval status value boosting for some criteria or fields. After judging, for almost half of the topics, a minimum of 50% of the documents retrieved are relevant, up to 90% for 10 of the 38 topics provided. Almost two thirds of the top-10 returned documents are judged relevant, while nearly a quarter of the relevant treatments is returned in the top-10 abstracts. Our best runs achieve highly competitive results depending on the measures, with on average being ranked #2 or #3 according to the official results for the literature task.

Introduction

The SIB Text Mining group [1], at the Swiss Institute of Bioinformatics (SIB) in Geneva, has been participating in several TREC campaigns: TREC Medical Records [2], TREC Clinical Decision Support [3,4], TREC Genomics [5], TREC Chemical IR [6] tracks and both TREC 2017 and 2018 Precision Medicine track [7,8].

In the context of national personalized health initiative, so-called SPHN (Swiss Personalized Health Network), the group is involved in a research project named “Swiss Variant Interpretation Platform for Oncology” (SVIP-O) [9]. A number of Swiss hospitals, pathology institutes and the Swiss Institute of Bioinformatics members have jointed their expertise to support the harmonization of variant annotation in diagnosis and to provide a centralized curated database for somatic variants from Swiss hospitals. In this context, the SIB Text Mining group developed a set of text analytics services intended to facilitate and improve the comprehensive collection of literature relative to a particular case and to prioritize variants (e.g. from VCF files) [10].

As in 2017 and 2018, the TREC 2019 Precision Medicine track focuses on the identification of scientific articles and clinical trials of interest for clinicians treating patients with cancer. The general structure and evaluation of TREC Precision Medicine 2019 are very similar to those of previous years. The main change lies in the possibility to return a set of treatments for each scientific abstracts, in order to evaluate the ability of the system to favor abstracts that are discussing various treatment options.

The scientific abstracts retrieval is based on two steps: the collection of the most comprehensive set of abstracts and the ranking of this set. To gather the set of abstracts, we first retrieve abstracts mentioning all entities of the triplet (i.e. the disease, the gene and the variant). Then we test different options to extend this set: relaxing the constraints (e.g. searching for abstract mentioning the disease and the gene, but not the variant) and/or using full-text articles to increase the chance to match the whole triplet. To rank the final set of abstracts, strategies successfully tested last year were again applied (e.g. re-ranking based on drug density). In addition, and to answer the new request of this task, we investigated two strategies to enable a more diverse set of treatments to be present in the top returned abstracts.

For the clinical trial task, the first strategy performed a query with a maximum of constraints about demographic data, disease and gene. While variants are not widely indicated in clinical trials, it would have been too restrictive to force their presence in the query. However, to favorize clinical trials containing the required variant, a boost on the ranking score was applied. For further strategies, constraints were gradually relaxed, using linear combination of queries, disease expansion to retrieve more general/specific diseases, generation of variant synonyms, etc. Based on the experience of last year, showing that usage of additional textual fields for the retrieval of diseases enables better performance, we applied this strategy for the retrieval of all entities this year.

1. Data

The Precision Medicine track provides two collections, one for each task: scientific abstracts and clinical trials. Both tasks share a common topics set.

1.1 Scientific abstracts

While the scientific abstracts collection was composed in 2017 and 2018 of a snapshot of PubMed abstracts and additional abstracts from oncology conferences' proceedings (i.e. AACR and ASCO proceedings), the collection evolved in 2019. It is now exclusively composed of PubMed abstracts. The 2019 snapshot covers PubMed abstracts published until mid-December 2018, corresponding to 28,137,808 abstracts.

1.2 Clinical trials

The 2019 collection is composed of 306,238 documents in XML format, and regroups clinical trials from NCT00000102 submitted in 1999 to NCT03958253 submitted in May 2019. In these XML files, information about clinical trials is available such as title, summary, description, status, etc. For some sections of a clinical trial, the information is well structured such as the location, gender or age of the patients. Nevertheless, some sections remain unstructured (i.e. free text) like inclusion or exclusion criteria.

1.3 Topics

This year, the topics set is composed of 40 semi-structured synthetic cases, created by precision oncologists at the University of Texas MD Anderson Cancer Center. Similarly as last year topics, each topic consists of three fields: the *disease* represents the diagnosis, the *gene* contains the genetic variant and the *demographic* field mentions the age and gender of the patient. Regarding the *gene* field, the type of content differs among topics. Indeed, 14 topics mention only a gene without specification of the genetic variants, 14 topics contain a gene with a single nucleotide variant (e.g. BRAF (E586K)), while the other 12 topics contain other types of variations, such as copy number variants (e.g. ERBB2 amplification), genes fusion (e.g. RANBP2-ALK fusion) or combination of two variants, etc. It is also to be noted that in one case, the gene is not mentioned (i.e. topic 15 with high tumor mutational burden).

1.4 Ontologies and Resources

Several publicly available ontologies and resources have been used for developing our systems.

Gene - neXtProt. Developed by the SIB (Swiss Institute of Bioinformatics) in 2008, the neXtProt human protein knowledgebase [11] is a comprehensive human-centric discovery platform. More than 20,000 proteins were manually annotated and still updated. This provides researchers with high-quality synonyms for both protein and gene names. On our side, for the Precision Medicine track, we used neXtProt to normalize gene names.

NCI Thesaurus. We used the NCI Thesaurus (NCIt) [12] for disease mapping. It covers clinical care, translational and basic research, public information and administrative activities. Provided by the National Cancer Institute, this terminology is a standard for biomedical coding and reference, used both by public and private scientific partners worldwide. This NCI's reference terminology contains the NCI_CUI, the semantic type, a preferred term, some NCI and MeSH synonyms.

Drugbank. The Drugbank database [13] is a freely accessible resource which includes more than 13,000 records (version 5.1.4, released 2019-07-20). It contains information on drugs (i.e. pharmacological, chemical and pharmaceutical) and drug targets (i.e. structure, sequence, pathway), synonyms and product names.

MeSH. Provided by the U.S. National Library of Medicine (NLM), the Medical Subject Headings (MeSH) [14] is a controlled vocabulary thesaurus used for indexing articles for PubMed. In comparison with specialized ontologies like the NCIt, MeSH is less granular and easily identified by Natural Language Processing thanks to synonyms.

Variant synonym generation tool. This tool [10,15], developed by our group, generates synonyms with many syntactic variations encountered in the literature [16], as well as description of the variant at other levels (e.g. genomic level) using the Mutalyzer tool [17]. It generates up to 42 synonyms for a given variant, with 16 synonyms for protein variant, 13 for transcript variant including COSMIC identifier and 13 for genomic variant including dbSNP identifier. For instance, *Val143Ala* and *428T>C* are two of the synonyms generated for the *V143A* variant.

2. Strategies

2.1 Collection pre-processing

A pre-processing of both collections (i.e. the MEDLINE and clinical trials collections) was performed before query time. For each collection, fields of interest were selected. The title, abstract, MeSH terms, chemicals and keywords were chosen for the MEDLINE collection. The titles (brief and official), inclusion and exclusion criteria, brief summary, detailed description, condition, age (minimum and maximum) and keywords were selected for the clinical trials collection. These fields were extracted from the original documents and loaded into a MongoDB document database.

Then, the fields were annotated with codes from various terminologies, i.e. neXtProt for genes, DrugBank for drugs, NCI Thesaurus for diseases. As an example, if the word "BRAF" is mentioned in an abstract, the annotation "NX_P15056" will be created for the relative abstract. In the same way, if the description field of a clinical trial mentions "melanoma", a new field named "description_annotated" will be created and contains "C3224" as an annotation. Querying our collections through annotations is not only faster because the indexes are pre-computed, it also results in a better recall as each occurrence of a concept receives one unique identifier for all its synonyms. In addition, the annotation process implies string pre-processing methods. For instance, when a dash is present in a term from our terminology, it is transformed to a set of additional words (e.g. “AB-C” is transformed to “AB”, “C” and “ABC”). Such processing enables us to retrieve papers in which only occurrences of the word with the dash are present. The annotations are then pushed into the MongoDB database.

No annotation has been performed for the variants because of a serious limitation: no universal terminology is available. However, original fields are also stored to search variants in free text.

Finally, each document (i.e. a MEDLINE abstract or a clinical trial) were indexed in an ElasticSearch index (version 7.2.0).

2.2 Scientific abstracts retrieval

For our third participation to the scientific abstracts task, we reused successful strategies applied previous years (e.g. re-ranking based on drug occurrences, queries with decreasing level of specificity). The topics and relevance judgments of 2018 have been used as training data to select the optimal tuning for each strategy. Our system is based on two steps: 1) collection of a complete set of abstracts and 2) ranking of this set of abstract.

In addition, this year, we had the possibility to return up to three treatments per abstract. The returned treatment must be a string corresponding to the exact text found in the abstract. To this extent, we used our drugs annotations (i.e. DrugBank annotations) of the MEDLINE collection. The top-3 drugs annotations are selected using the following rules: if more than three drugs have been found in the abstract, drugs related to oncology are selected. If the same drug appears with different terms in the abstract, the DrugBank preferred term is chosen.

2.2.1 Collection of a complete set of abstracts

The collection of a complete set of abstracts related to a particular triplet (i.e. a variant in a gene for a specific diagnosis) is built by the union of three queries (Figure 1). The first query is the intersection of two queries’ output: the gene and the disease are searched with unique identifiers within the annotations, while a keyword search expanded with synonyms is performed within free text abstracts for the variant. The second query is the intersection of two querie’s output: the disease is searched with a unique identifier within the annotations and the gene and variant are merged together to form a new word (e.g. BRAF and V600E becomes BRAFV600E), which is searched within free text abstracts. The third query searches each exact topic term within free text abstracts. When several genes and/or variants are present in the topic, we treat each of them as a subtopic and merge the set of abstracts retrieved for each subtopic to define the final set of abstracts for the topic. When no variant or not gene is present in the topic, only the first and third queries are performed.

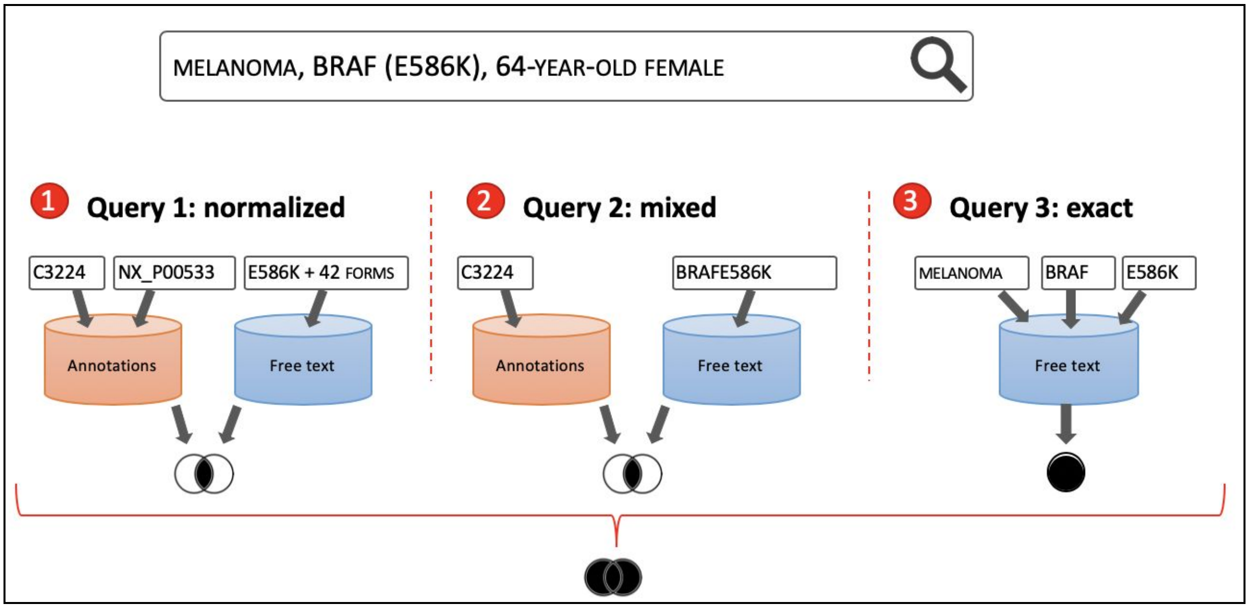

Figure 1. Example of the three queries generated for the topic 1

On top of that, several strategies have been tested to expand the set of abstracts. Indeed, while MEDLINE content is usually sufficient to decide whether a particular report needs in-depth reading or not, it is also to be noted that scientific abstracts reporting on treatments do not always mention all the information regarding disease, gene and variant. Moreover, an abstract about the given variant but for another cancer type may still be valuable from a clinical point of view. Therefore, our first strategy aims at collecting abstracts with decreasing levels of specificity: abstracts not mentioning one of the entities of the triplet (e.g. abstracts not mentioning the disease). In this first strategy, results are linearly combined with previous abstracts set, with different weights for each expanded query. The second strategy collects abstracts using the same approach, but combines them in a different manner: exact results are returned first, and the results obtained through less specific queries are combined together and returned afterwards. The third strategy expands the abstract set using full text articles. A local collection of PubMed Central (PMC) is used. Diseases and genes are searched within PMC annotations while variants are searched in the full text articles. PMIDs corresponding to the query results are retrieved and linearly combined with the MEDLINE abstracts set.

2.2.2 Ranking of the abstracts set

The ranking of the abstracts set includes several strategies to re-rank the MEDLINE set. A common core of three re-ranking strategies has been applied to all runs, while two additional strategies have been tested for some specific runs. These two additional strategies have been developed to answer the new subtask consisting to avoid that all top-returned abstracts are related to the same treatment, but rather to favor abstracts that are discussing different treatment options.

First, the score, initially based on the relevance score provided by Elasticsearch, is recalculated based on the number of occurrences of some annotated entities (i.e. gene, disease and drug). It is based on the assumption that an abstract with a high frequency of these entities is probably more relevant to support our task. Annotations of the collection are used to calculate the density of an entity (i.e. the number of occurrences of a concept) in the abstract, title, MeSH terms, chemicals and keywords. In addition, for the drugs, a cancer-related density is calculated using a list of 633 DrugBank records. This list has been manually created using different resources: cancer-related categories from DrugBank (e.g. *Antineoplastic agents*), the Cancer Drugs List from the National Cancer Institute [18], the List of Cancer Chemotherapy Drugs from the Navigating Care [19] and the Oral Chemotherapy Drugs List from CareFirst [20]. Different weights are attributed depending on the entity types.

Second, the score is recomputed using demographic information. This strategy follows last year's experiments. Age-groups and genders are extracted from the MeSH terms associated with MEDLINE abstracts. When the abstract is targeting the required gender and/or age-group, a strong positive boost is attributed. If the gender and/or the age-group is not discussed in the abstract, the abstract might still be relevant, therefore, it also obtains a positive boost, but smaller. Finally, when the gender and/or the age-group is not matching the required one(s), the abstract obtains a negative penalty.

Third, we propose a re-ranking aiming to increase the relevance of abstracts related to precision medicine. To this extent, we reuse the lists of positive keywords (e.g. “treat”) and negative keywords (e.g. “marker”) developed for the TREC PM 2017. These lists contain stemmed keywords, based on a manual screening of a subset of retrieved articles, as well as the testing of list variations on the basic tuning set. The presence of positive keywords in an abstract will result in a better scoring, while the negative keywords will decrease its relevance.

The first treatments re-ranking strategy tested to improve the breadth of treatments returned in the top-positions selects each drug appearing in the results. For each drug, the list of abstracts mentioning this drug is collected. Thus, an abstract containing different drugs is present in several lists. Scores are then normalized so that each top-result for each drug gets a score of 1. Finally, the lists of results for each drug are merged together. A new score is then calculated for each abstract. The second treatments re-ranking strategy tested relies on the ranked abstracts set. An abstract that provides a not-already seen drugs is kept in its top-position, while the score of an abstract with already mentioned drugs is pushed after the last abstract bringing novel drug information.

2.2.3 Runs

Five runs have been submitted (Table 1).

Strategies	SIBTMlit1	SIBTMlit2	SIBTMlit3	SIBTMlit4	SIBTMlit5
Collection of abstracts					
Exact results	X	X	X	X	X
Expanded results linearly combined	X	X	X		X
Expanded results pushed afterwards				X	
PMC results linearly combined					X
Ranking strategies					
Entities density	X	X	X	X	X
Demographic information	X	X	X	X	X
Lists of keywords	X	X	X	X	X
First treatments re-ranking strategy		X		X	X
Second treatments re-ranking strategy			X		

Table 1: Resume of the strategies tested for the five submitted runs

Our first run (SIBTMlit1) is based on the abstracts set generated by the linear combination of the exact results (Disease + Gene + Variant) together with the expanded results (Disease + Gene; Disease + Variant; Gene + Variant). The re-ranking is using the common core set of strategies: based on entities density, on demographic information and on the presence of the predetermined positive and negative keywords.

Our second run (SIBTMlit2) repeats the strategy of the run SIBTMlit1. In addition, the first strategy to obtain a large variety of treatments in the top-positions is used.

Our third run (SIBTMlit3) also repeats strategy of the run SIBTM1, but including this time the second strategy to obtain different treatments in the top-positions.

Our fourth run (SIBTMlit4) is based on the collection of exact results (Disease + Gene + Variant), in which expanded results are pushed afterwards (Disease + Gene; Disease + Variant; Gene + Variant). Similarly, the common core set of re-ranking strategies are used. Finally, the first treatment re-ranking strategy is applied.

Our fifth run (SIBTMlit5) is based on the linear combination of the exact results (Disease + Gene + Variant), together with the expanded results (Disease + Gene; Disease + Variant; Gene + Variant), as well as the PMC results. The three common re-ranking strategies are then applied, as well as the first treatment re-ranking strategy.

2.3 Clinical trials retrieval

Based on the Elasticsearch index built in the pre-processing step, several strategies have been tested to select clinical trials of interest for topics' patients.

2.3.1 Description of queries

Four queries are designed as presented in Table 2. Differences between them are about constraints relaxing and query expansions. For gene and variant, expansion is done through regexes and synonyms. For disease, a supplementary degree is added to take into account parent/children diseases, for more general or more precise pathologies respectively.

For Query 1, demographic data (gender and age) match with the topic or are not discussed. Gender has to be identical to topic's gender or equals “all”, age has to be equal or superior to the minimal age and inferior to the maximal age indicated in the clinical trial. Relevant codes for gene and disease have to be present in one of the annotated fields of the clinical trial.

Query 2 differs from the first query by an expansion of the disease. Thanks to NCI Thesaurus*,* a document that compiles disease's relations was created. For each disease code, a direct parent disease and a list of parent/children diseases more general or specific are associated (e.g. [general] *cancer* < *carcinoma* < *lung carcinoma* [specific]). A clinical trial considered irrelevant regarding the exact disease code will be retrieved if the code of a parental disease is founded.

Query 3 is nearly identical to Query 2. The only difference is the expansion of variant thanks to regexes usage and synonyms provided by the variant synonyms service described in section 1.4.

Query 4 was an additional query, not submitted alone, exclusively in combination with query 2 and query 3 for run 3 and run 5 respectively. This fourth query allows the absence of the gene in clinical trials but the presence of the variant is required to maintain specificity.

Queries	Query 1	Query 2	Query 3	Query 4
Query parameters				
Demographic data				
Exact or not discussed	X	X	X	X
Disease				
Exact	X	X	X	X
Expansion to more general diseases		X	X	X
Expansion to more specific diseases		X	X	X
Gene				
Exact	X	X	X	
Regex expansion		X	X	
Variant				
Exact or not discussed	X	X	X	X
Regex & synonyms expansion			X	X
Boost				
Major boost if variant = exact	X	X	X	X
Depending fields		X	X	X
Depending parent level disease		X	X	X

Table 2: Resume of the strategies tested through 4 queries.

2.3.2 Runs

Five runs have been submitted for the clinical trials task. Table 3 presents the correspondence between each query and each run, as well as the combination when applicable.

Runs	SIBTMct1	SIBTMct2	SIBTMct3	SIBTMct4	SIBTMct5
Query 1	X				
Query 2		X	X		
Query 3				X	X
Query 4			X		X

Table 3: Correspondence between queries and runs submitted for the Clinical Trials task.

SIBTMct1 is the most constraining run. Composed by the query 1, all criteria (expected the variant) have to be present in the clinical trial in order to be selected.

In the SIBTMct2, the first constraint relaxing concerns expansions on disease and gene criteria as well as boost on fields and hierarchic level diseases.

The SIBTMct3 is based on the same query as SIBTMct2. The shade lies in the linear combination of this query 2 with the query 4, where constraints are “inverted”, more relaxing about gene but more restrictive about variant.

In the SIBTMct4, an additional level of constraints relaxing is added. This time, expansion of variants (synonyms and regex) is taken into account. Boosts application remains the same.

SIBTMct5, like SIBTMct3, is based on the combination of two queries. For this run, the combination is made up of query 3 and query 4.

3. Results and Discussion

In this section, we present the results for the scientific abstracts retrieval task and the clinical trials retrieval task.

Metrics used to evaluate the document ranking are infNDCG, P@10 and R-Prec. The infNDCG (inferred non discounted cumulative gain) reflects the gain brought by a document based on its position in the ranked results. P@10 (precision at rank 10) represents the proportion of documents retrieved in the top ten results that are relevant. It thus reflects the ability of the system to retrieve relevant results at high ranks. Finally, R-Prec (R-Precision) returns the number of relevant documents returned in the top R document, where R corresponds to the number of relevant documents for the query. Metrics used to evaluate the treatment ranking are R@10 and F1@10. The R@10 (recall at position 10) represents the proportion of relevant treatments retrieved among the top 10 retrieved results. The F1 (also called F-measure) is the harmonic mean of the recall and the precision.

3.1 Scientific abstracts retrieval

3.1.1 Tuning Results

The topics and relevance judgements of TREC PM 2018 have been used to select the best settings for our system.

When a linear combination of exact results with expanded results was performed, scores from the exact query receive a weight of 1.0, scores of the “Disease+Gene” query receives a weight of 0.96, scores of the “Disease+Variant” query receives a weight of 0.05 and scores of the “Gene+Variant” query receives a weight of 0.07. When PMC results are combined with MEDLINE results, a weight of 0.9 is given to MEDLINE scores and a weight of 0.1 to PMC scores.

The entity-related density re-ranking performed the best when a weight of 0.14 was given to drug density, a weight of 0.76 to cancer-related drug density, a weight of 0.75 to disease density and a weight of 0.7 to gene density. The age-related re-ranking showed its best performance when we applied a positive boost of respectively 1.0 and 0.3 to abstracts mentioning the requested age-group and to abstracts containing no age-group information and a negative boost of 0.05 to abstracts targeting another age-groups. Similarly, the gender-related re-ranking performed the best with positive boosts of 0.8 and 0.6 for abstracts targeting the requested gender and abstracts not mentioning gender, and a negative boost of 0.05 for abstracts about a different gender. The keywords re-ranking had the most impact on our results when using a positive boost of 0.4 for positive keywords a negative penalty of 0.15 for negative keywords. The selected positive keywords list is composed of the following stemmed keywords: *treat;drug;therap;prognos;surviv*. The optimal negative keywords list is composed of the following stemmed keywords: *immuno;marker;detect;sequencing*.

Finally, the final score of an abstract was composed by the MEDLINE score (0.5), the entity-related re-ranking (0.07), the demographic-related re-ranking (0.18) and the keywords-related re-ranking (0.23).

3.1.2 Final Results

Results for the 40 topics are presented in Table 4. The baseline system (SIBTMlit1) performs the best regarding precision at rank 10: 62.8% of the abstracts returned in the top-10 positions are judged as relevant. Re-ranking the abstracts list using our first strategy for increasing the diversity of treatments returned in the top-positions shows a positive impact regarding R-Precision: the SIBTMlit2 reaches a R-Prec of 31.7%. While the re-ranking using the second strategy results in a strong decrease of our results regarding the abstracts ranking, it strongly improves our results regarding the treatments extraction. The SIBTMlit3 obtains a R@10 of 24.1% and a F1@10 at rank 10 of 28.2%. Finally, the use of full-texts to re-rank our results improves the infNDCG, which reaches 53.4% for SIBTMlit5.

	Abstracts ranking			**Treatments ranking**	
	infNDCG	**P@10**	**R-Prec**	**R@10**	**F1@10**
SIBTMlit1	0.531	**0.628**	0.310	0.098	0.131
SIBTMlit2	0.531	0.623	**0.317**	0.153	0.199
SIBTMlit3	0.391	0.558	0.205	**0.241**	**0.282**
SIBTMlit4	0.531	0.623	**0.317**	0.153	0.199
SIBTMlit5	**0.534**	0.615	0.312	0.150	0.191

Table 4: Resume of the results for the five submitted runs for the scientific abstracts task.

Almost a quarter of the relevant treatments are returned in the top-10 abstracts. However, the evaluation of this new task faces a strong limitation: treatments are returned as non-normalized entities (i.e. uncased strings).

An overview of the P@10 topic per topic is presented in Figure 2. Queries relative to single nucleotide variants obtain encouraging results: for our best run regarding P@10 (SIBTMlit1), each query concerning a single nucleotide variant (SNV) obtains at least 60% of P@10. However, queries relative to gene fusion do not perform well: these queries obtain between 10% and 40% of P@10. No specific trend could be deducted for queries with no mention of genetic variants.

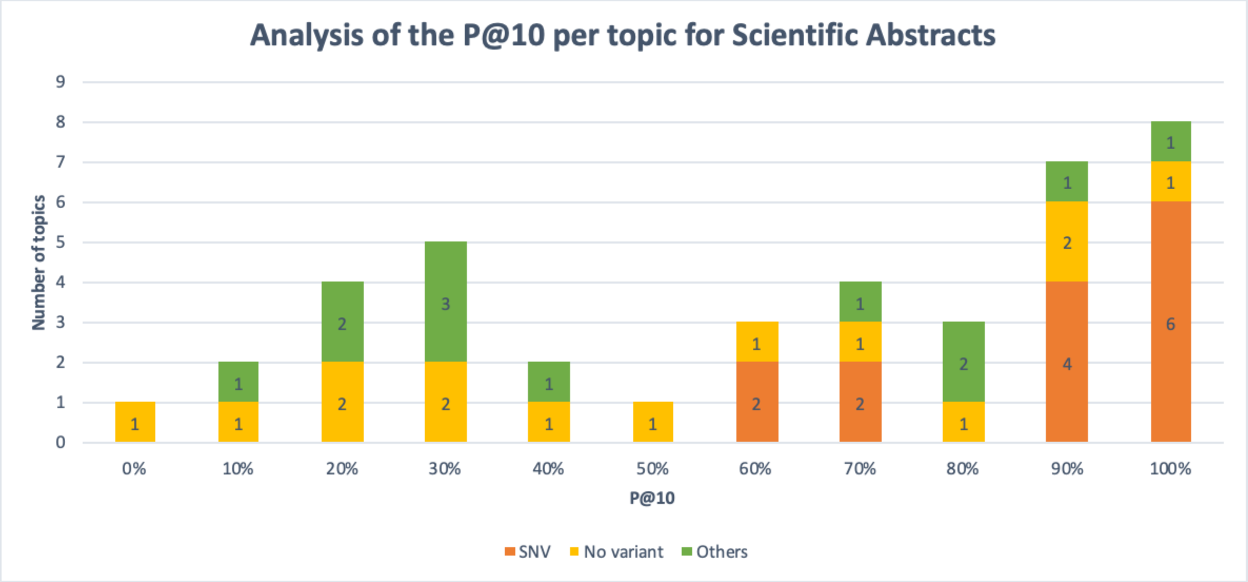

Figure 2. Distribution of topics (number and type) according to the P@10 values obtained for the scientific abstracts task.

Our system is composed of two steps: collecting a set of abstracts and re-ranking these abstracts. While the second step is important to prioritize the literature, the first step is mandatory: the relevant abstracts must be captured in our set to be properly ranked. We thus investigate which proportions of relevant abstracts have been successfully retrieved by our system, whatever ranks they were returned. For almost half of the topics, more than 70% of the relevant documents are retrieved in our abstract set. For such topics, improving the ranking could be beneficial. However, for nine topics, less than half of the relevant documents are retrieved. Further investigations are needed to analyze such abstracts and define possible actions to improve their gathering, such as improving the annotations of genes or diseases, expanding the diseases to parent and/or children diseases, defining better synonym lists for genetic variants, etc. No specific trend could be deducted regarding types of topics.

3.2 Clinical trials retrieval

3.2.1 Tuning Results

Similarly to the scientific abstracts task, topics and qrels files of TREC PM 2018 have been used to select the best settings in case of boost or linear combination.

There were three kinds of boost: 1) regarding the presence of variant; 2) relating to the hierarchy level of diseases listing; 3) about the importance to be given to the different fields of the clinical trial. Firstly, boost about variant was defined arbitrarily, a big value was applied to be sure that clinical trial with variant (even if cited only once) was higher than the others. In a second time, boosts about hierarchical level diseases were tuned and resulted in 5 for an exact disease found, 1 for a direct parent disease, 0.4 for one of the children diseases, 0.1 for one of the parent diseases. Finally, another boost was tuned, relating to fields in which information was retrieved: 1 if finding in the title, 0.3 if in the official title, 1 if in the summary, 1 if in the detailed description, 1.15 if in condition, 1 if in criteria, 1 if in inclusion criteria, 1 if in keywords.

Many weight combinations for the linear combination of SIBTMct3 and SIBTMct5 runs were tested. The best combination was to multiply by 0.6 the score of classical queries (query 2 or 3) and by 0.4 the one of the extra query 4.

3.2.2 Final Results

As explained and described on the top of the results part, metrics used to evaluate our results are infNDCG, P@10 and R-Prec. For the clinical trials task, evaluation is based on 38 topics. Indeed, two topics, 32 and 33, were judged out of scope (no relevant documents have been found in clinical trials collection) so they were dropped for the evaluation.

Averages of these metrics for each run (SIBTMct1 to SIBTMct5) are shown in Table 5.

	infNDCG	**P@10**	**R-Prec**
SIBTMct1	0.374	0.382	0.249
SIBTMct2	**0.496**	0.458	0.366
SIBTMct3	0.487	**0.471**	0.362
SIBTMct4	**0.496**	0.468	**0.370**
SIBTMct5	0.474	0.463	0.353

Table 5: Resume of the results for the five submitted runs for the clinical task.

As shown, the baseline run (SIBTMct1) displays the worst metrics. This confirms that the release assumptions and boosts applied to the following runs improve the clinical trials retrieval. The SIBTMct3 obtains the best P@10 value (0.471) but we could easily affirm that the best run is SIBTMct4. Indeed, it shows the best infNDCG and R-Prec metrics and P@10 value is not really much lower (0.468 vs 0.471). The addition of constraints relaxing in our fifth strategy does not induce better scores than SIBTMct4.This reflects that reducing considerably the constraint on the gene is not wise, since the deficit of specificity is not compensated by forcing the presence of variant.

Figure 3 displays an overview of the P@10 topic per topic for the best run (SIBTMct4). As for scientific abstracts task, frequency of single nucleotide variants (SNV) queries in the best results is higher than other kinds of queries. On the 15 topics related to SNV, 60% of these specific topics obtain a P@10 ≥ 90%, which means that 90% of the top-ten documents retrieved by the system are judged relevant. On the other hand, in the case of unspecified variant or “other” (that includes terms like “fusion”), the trend is more in the low range with P@10 ≤ 50%.

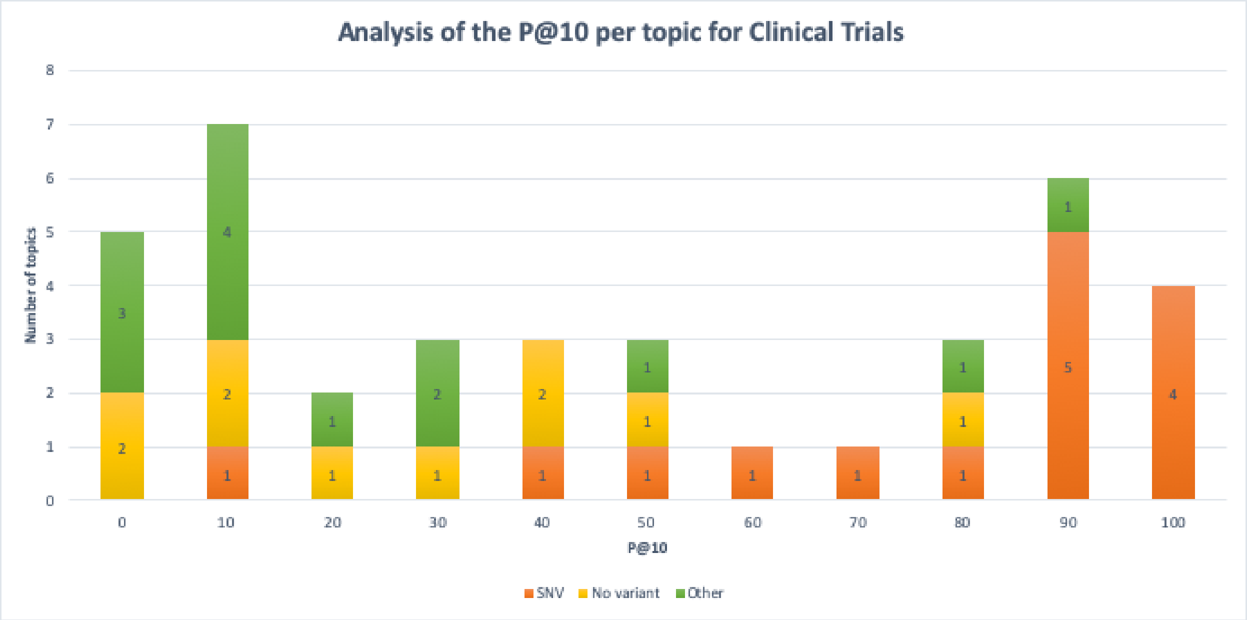

Figure 3. Distribution of topics (number and type) according to the P@10 values obtained for the Clinical Trials task.

3.3 Impact of variant synonyms

Our variants synonyms generation tool combines two approaches: syntactic variations and levels of description (genome, genes and proteins). However, from the TREC evaluation guidelines, it is unclear if sequence variants expressed using nucleotides (genes and genome levels) are taken into account by the assessors. To try to answer this question, we can compare run 2 (SIBTMct2) and run 4 (SIBTMct4) of the clinical trials task that only differ by the use of variants synonyms (Table 6).

	Run 2 (SIBTMct2)	Run 4 (SIBTMct4)
infNDCG	0.496	0.496
P@10	0.458	0.468
R-Prec	0.366	0.370

Table 6: Comparison of run 2 and run 4 for the clinical trials.

We observe a marginal impact of the synonyms. It may be due to the detection of syntactic synonyms rather than nucleotide synonyms, or to the detection of nucleotide synonyms in clinical trials already mentioning the variant in amino acids (protein level).. While we cannot assess here the extent of potential variants synonyms overlook, table 7 indicates a high impact of variants synonyms, with a recall gain from 0 to 467% (+127% on average), especially for VUS (Variant of Unknown Significance) and variants with insufficient functional characterization. For such variants, the ability to obtain more (relevant) articles at curation time is critical.

Gene	Variant	Medline	Expansion	Recall gain
GNAQ	Q209L	20	20	-
R-Prec	L858R	1323	1341	+1%
PIK3CA	E545K	179	191	+7%
BRAF	V600E	3859	4293	+11%
BRCA1	P871L	13	26	+100%
FLT3	N676	3	12	+300%
TP53	V143A	6	34	+467%

Table 7: Recall gain obtained with variants synonyms expansion.

Conclusion

For this TREC edition, our results were ranked at the 2nd or 3rd position for the literature task and between the 5th and 7th positions for the clinical trials task, depending on the measures.

Despite two distinct approaches for the two tasks, we noticed some common trends among results. In both cases, our systems performed the best on the topics concerning single nucleotide variants (e.g. topics number 6 and 7), while topics about genes fusion (e.g. topics number 13 and 18) resulted in poor performances. Such results suggest that our approach to capture variants using a synonyms generation tool is efficient to improve the retrieval and ranking of relevant documents. Indeed, our tool is currently focused on expanding single nucleotide variants.

However, the usage of variant synonyms for the TREC PM track needs further investigations. Indeed, based on the evaluation's methodology presented during the TREC 2019 conference, documents mentioning the variant using only a synonym term are supposed to be rejected. Exploring the extent of variants synonyms overlook and its impact may be of importance given its critical effect for clinical interpretation of poorly characterized variants.

Acknowledgments

This experiment has been supported by the Swiss Personalized Health Network (SPHN) and BioMedIT fundings (see https://svip.ch/) and partially supported by the MyHealthMyData project

#732907. This work also benefited from discussions with Daniel Stekhoven, Valérie Barbié and other members of the SVIP-O project.

References

[1] "BiTeM." [Online]. Available: http://bitem.hesge.ch/

[2] J Gobeill, A Gaudinat, E Pasche, D Teodoro, D Vishnyakova, and P Ruch. BiTeM Group Report for TREC Medical Records Track 2011. In TREC. 2011.

[3] J Gobeill, A Gaudinat, E Pasche, and P Ruch. Full-texts representation with Medical Subject Headings and co-citations network reranking strategies for TREC 2014 Clinical Decision Support Track. In TREC. 2014.

[4] J Gobeill, A Gaudinat, and P Ruch. Exploiting incoming and outgoing citations for improving Information Retrieval in the TREC 2015 Clinical Decision Support Track. In TREC. 2015.

[5] J Gobeill, F Ehrler, I Tbahriti, and P Ruch. Vocabulary-driven Passage Retrieval for Question-Answering in Genomics. In TREC. 2007.

[6] J Gobeill, A Gaudinat, E Pasche, D Teodoro, D Vishnyakova, and P Ruch. BiTeM group report for TREC Chemical IR Track 2011. In TREC. 2011.

[7] E Pasche, J Gobeill, L Mottin, A Mottaz, D Teodoro, P Van Rijen, and P Ruch. Customizing a Variant Annotation-Support Tool: an Inquiry into Probability Ranking Principles for TREC Precision Medicine. In TREC. 2017.

[8] E Pasche, J Gobeill, L Mottin, A Mottaz, D Teodoro, P Van Rijen, and P Ruch. SIB Text Mining at TREC 2018 Precision Medicine Track. In TREC. 2018.

[9] DJ Stekhoven, P Ruch and V Barbié. Swiss Variant Interpretation Platform for Oncology (SVIP-O), Swiss Med Informatics 34 (2018), 00411.

[10] D Caucheteur, J Gobeill, A Mottaz, E Pasche, PA Michel, L Mottin DJ Stekhoven, V Barbié and P Ruch. Text-mining Services of the Swiss Variant Interpretation Platform for Oncology. MIE 2020 (accepted)

[11] P. Gaudet, P.A. Michel, M. Zahn-Zabal, et al. The neXtProt knowledgebase on human proteins: 2017 update, Nucleic Acids Res 45(D1) (2017), D177-D182.

[12] N Sioutos, S de Coronado, HW Haber, et al. NCI Thesaurus: a semantic model integrating cancer related clinical and molecular information, J Biomed Inform 40(1) (2007), 30-43.

[13] DS Wishart, YD Feunang, AC Guo, et al. DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res 46(D1) (2018), D1074-D1082.

[14] F Minguet, L Van Den Boogerd, TM Salgado, C Correr, and F Fernandez-Llimos. Characterization of the Medical Subject Headings thesaurus for pharmacy. Am. J. Health. Syst. Pharm. 2014:71:1965-72.

[15] SynVar tool: http://goldorak.hesge.ch/synvar/

[16] YL Yip, N Lachenal, V Pillet, et al. Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase, J Bioinform Comput Biol 5(6) (2007), 1215-31.

[17] M Wildeman, E van Ophuizen, JT den Dunnen, et al. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker, Hum Mutat 29(1) (2008), 6-13.

[18] "A to Z List of Cancer Drugs", National Cancer Institute". [Online] Available: https://www.cancer.gov/about-cancer/treatment/drugs. [Accessed: 01-Aug-2019].

[19] "List of Cancer Chemotherapy Drugs – Navigating Care". [Online]. Available: https://www.navigatingcare.com/library/all/chemotherapy_dru gs. [Accessed: 01-Aug-2019].

[20] "CareFirst. Oral Chemotherapy Drugs". [Online]. Available: https://member.carefirst.com/carefirst-resources/pdf/oral-chemotherapy-drug-list-sum2714.pdf. [Accessed: 01-Aug-2019].