Customizing a Variant Annotation-Support Tool: an Inquiry into Probability Ranking Principles for TREC Precision Medicine

Emilie Pasche[a,b], Julien Gobeill[a,b], Luc Mottin[a,b], Anaïs Mottaz[a,b,c], Douglas Teodoro[a,b], Paul Van Rijen[a] , Patrick Ruch[a,b]

[a] *HES-SO / HEG Geneva, Information Sciences, Geneva, Switzerland*
[b] *SIB Text Mining, Swiss Institute of Bioinformatics, Geneva, Switzerland*
[c] *Laboratory of Cognitive Neurorehabilitation, Faculty of Medicine, University of Geneva, Switzerland*
contact: {emilie.pasche;julien.gobeill}@hesge.ch

Abstract

The TREC 2017 Precision Medicine Track aims at building systems providing meaningful precision medicine-related information to clinicians in the field of oncology. The track includes two tasks: 1) retrieving scientific abstracts addressing treatment effect and prognosis of a disease and 2) retrieving clinical trials for which a patient is eligible. The SIB Text Mining group participated in both tasks. Regarding the retrieval of scientific abstracts, we designed a set of different queries with decreasing levels of specificity. The idea was to start initiating a very specific query, from which less specific queries will be inferred. We may thus consider as relevant abstracts that did not mention all critical aspects of the complete query but could still be of interest. Therefore, the main component of our approach was a large query generation module (e.g. disease + gene + variant; disease + gene; gene + variant) – with each generated query being differentially weighted. To increase the scope of the queries, we applied query expansion strategies. In particular, a single nucleotide variant (SNV) generator was developed to recognize standard nomenclature as described by the Human Genome Variation Society (HGVS) as well as non-standard formats frequently found in the literature. We thus expect to retrieve a maximum of relevant abstracts. We then applied different strategies to favor relevant abstracts by re-ranking them based on more general criteria. First, we assumed that an abstract with a high frequency of drug names is more probably relevant to support our task. Therefore, we pre-annotated all the collection with DrugBank, thus enabling to retrieve the number of occurrences of drug names per abstract. Second, we assumed that the presence of some specific keywords (e.g. "treat") in the abstract should increase the relevance of the paper, while the presence of some other keywords (e.g. "marker") should decrease its relevance. Third, we assumed that some publications, such as clinical trials, should receive higher relevance for this task. Regarding the retrieval of clinical trials, we investigated for the competition different combinations of filtering and information retrieval strategies, mostly based on the exploitation of ontologies. Our preliminary analysis of the collection showed that : (1) demographic features (age and gender) are stored in a perfectly-structured form in clinical trials, thus this feature can be easily handled with strict filtering ; (2) the trials contain very few mentions of the requested genes and variants ; (3) diseases are stored in very inconsistent forms, as they are free text entities and can be mentioned in different fields such as condition, keywords, summary, criteria, etc. Thus, we assumed that identifying clinical trials dealing with the correct disease was the most challenging issue for this competition. For such a task, we perform Name Entity Recognition with the NCI thesaurus in order to recognize mentions of diseases in topics and in different fields of the clinical trials. This strategy handles several issues of free text descriptions, such as synonyms ("Cancer" and "Neoplasm" are equivalent) and hierarchies ("Colon carcinoma" is a subtype of "Colorectal carcinoma"). Then, for each topic, we apply different strategies of trials filtering – according to fields where the disease was identified – and hierarchies. Finally, classical information retrieval is performed with genes and variants as queries. The strictest filtering leads to an average of 62 retrieved trials per topic and tends to favor high precision, while the most relaxed filtering leads to an average of 379 retrieved trials per topic and tends to favor high recall. Yet, results show that the Precision values are poorly impacted by these strategies, while runs that favor Recall showed a better general behavior for this task.

Introduction

The SIB Text Mining group [1], at the Swiss Institute of Bioinformatics in Geneva, has a long history of participation in TREC campaigns, including TREC Genomics [2], TREC Medical Records [3], TREC Chemical IR [4] and TREC Clinical Decision Support [5, 6] tracks. In parallel, the group is currently involved in several translational medicine research projects, including the MD-Paedigree project (EU FP7 Programme), where its task was to help clinicians to retrieve similar cases in a federated digital repository gathering data from several European clinical centres, for

better personalized predictive medicine. Additionally, the group started to work on variants with the aim to provide text-mining tools to facilitate variant interpretation through literature mining (e.g. variants prioritization).

The TREC 2017 Precision Medicine Track focus on the identification of both scientific articles and clinical trials, regarded as useful for clinicians when treating patients in oncology. Generally, the topic consisted in: a disease, one or several mutated genes, some demographic information and some additional useful information, such as comorbidities. Two tasks were proposed. In the first task (i.e. the scientific abstracts retrieval task), participant's system had to return a ranked list of scientific abstracts considered as clinically useful from a precision-medicine point of view. In the second task (i.e. the clinical trials retrieval task), participant's system had to return a ranked list of clinical trials in which the patient may be (or would have been) eligible. The track provided no training data, and each group was allowed to submit up to five runs per task. The runs of both tasks were then evaluated by a pool of clinicians that judged the relevance of the submitted documents. Unfortunately, the guidelines used by these clinicians for judging relevance were not known at the submission time.

For producing the runs for the scientific abstracts task, we developed a core system, based on a set of queries, each differentially weighted using a curated data sample. Assuming this strategy would enable us to retrieve a large subset of relevant abstracts, we then applied different strategies to work on the ranking of the retrieved abstracts, such as using the number of occurrences of drug names or the publication types.

For producing the runs for the clinical trials task, we investigated and mixed different strategies based on Named Entity Recognition (NER), Information Retrieval (IR), and filtering, depending on the topic feature. On one hand the demographic feature is strongly structured in clinical trials, and was effectively managed by filtering, such as a selective query in a database. On the other hand, diseases are stored in very inconsistent forms in clinical trials, as they are free text entities and can be mentioned in different fields such as condition, keywords, summary, criteria, etc. We assume that identifying the possible conditions, and then applying more or less relaxed filtering, was the most challenging issue in this competition.

1. Data

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

1.1 Scientific abstracts

The scientific abstracts collection is composed of a snapshot of PubMed abstracts (January 2017) together with additional abstracts from AACR (American Association for Cancer Research) and ASCO (American Society of Clinical Ontology) proceedings. The XML version of the PubMed collection is used. It contains 26,670,000 abstracts, corresponding to 26,669,401 unique PMIDs. The latest version of a duplicated PMID is used. Title, abstract, publication date, publication types and MeSH terms are extracted for each abstract. AACR and ASCO abstracts are provided as TXT file. They contain respectively 33,018 and 37,007 abstracts. Only title, abstract and publication date are available for this subset.

1.2 Clinical trials

The scientific abstracts collection is composed of a snapshot of ClinicalTrials.gov (April 2017). It consisted of approximately 240,000 clinical trials in XML format. Clinical trials are semi-structured documents. They have dedicated sections for storing information, such as the study phase, the sponsors, the design of the study, or the eligibility criteria. Some sections contain formatted information (i.e. predefined values, such as demographic conditions) while others contain free text. Thus, different sections were exploited during the study.

1.3 Topics

The topics set consists of 30 semi-structured synthetic cases created by precision oncologists at the University of Texas MD Anderson Cancer Center. For each topic, the following information is mentioned: disease, one or several mutated gene(s) per case (including details about the variation in 26/37 cases), demographic data (i.e. age and sex) and additional information (e.g. comorbidities).

1.4 Ontologies and resources

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

UniProtKB [7] is developed in collaboration between the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). It is one of the main resources regarding protein and gene information. With 555,594 manually annotated records, UniProtKB provides high-quality synonyms for both protein and gene names.

NeXtProt is a protein-centric knowledgebase developed at the SIB Swiss Institute of Bioinformatics focused

solely on human proteins [8]. It includes an extended representation of human variants.

The dbSNP database is a National Center for Biotechnology Information (NCBI) resource that serves as a central, public repository for genetic variation [9]. dbSNP RefSNP cluster ID (rsid) are used to identify non-redundant sets of SNPs that map to an identical location.

The NCI Thesaurus (NCIt), provided by the National Cancer Institute, [10] is a reference terminology for biomedical coding, broadly used by both public and private care actors. This terminology covers clinical care, translational and basic research and public information and administrative activities. We used this resource for disease mapping, as it contains information for nearly 10,000 cancer and related diseases.

The Medical Subject Headings (MeSH) [11], provided by the US National Library of Medicine, is a controlled vocabulary used for indexing articles in MEDLINE. The MeSH is known for being less granular than specialized ontologies such as the NCIt, but also for being easily identified by Natural Language Processing, thanks to synonyms.

DrugBank [12] is a database containing biochemical and pharmacological information about drugs and drug targets. As of its latest release, DrugBank includes 10,500 records. A high number of synonyms are provided, as well as products names.

2. Strategies

In this section, we describe the strategies applied for each task.

2.1 Scientific abstracts retrieval

Our group has submitted five runs. As no official training data were provided for this task, we first built a basic training data set. We manually created 20 topics based on publicly resources such as ClinVar [13] and CIViC [14]. In average, there are 6.95 citations per case (minimum: 2; maximum: 31). However, this tuning set is limited by the facts that 1) ASCO and AACR articles are never cited; 2) few references are provided for each case and 3) no indication if the information of interest is located in the abstracts or in the full-text articles.

We then annotated diseases, genes and drugs within the whole collection, based on existing terminologies (i.e. NCIt for diseases, UniProtKB for genes and DrugBank for drugs). These annotations were then indexed together with the title, abstract, publication date, publication types and MeSH terms for each document. Solr Apache 6.6 is used for indexing and retrieval.

2.1.1 Baseline

The core of our system relies on a set of different queries with decreasing levels of specificity. Indeed, our assumption is based on the fact that an abstract of interest may sometimes not mention the specified variant, but for instance another variant affecting the gene in a similar manner. Similarly, an abstract about the variant of interest for another cancer type may still be valuable from a clinical point of view. Therefore, our approach is based on the generation of a set of three queries:

- Query 1: disease + gene + variant
- Query 2: disease + gene
- Query 3: gene + variant

Results are then merged together through linear combination. Each set of results is differentially weighted.

Moreover, we apply query expansion strategies. Synonyms of genes are generated using UniProtKB terminology, while synonyms of diseases are retrieved using NCIt. Regarding variants, a synonym list has been manually created for copy number variants (e.g. amplification), while a SNVs synonym generator has been developed for single nucleotide variants. Given variant information, that is gene name and amino acid change, the SNVs generator produced the standard nomenclature format at the protein level as described by the Human Genome Variation Society (HGVS) [15]. When a corresponding rsid was found through NeXtProt, the HGVS standard description was also generated for the different protein isoforms as well as for the transcript and genomic DNA description levels. Additionally, non-standard formats frequently found in the literature were generated for these different levels of description [16]. It included at the protein level the use of single and three letters amino acid codes (e.g. Val600Glu) as well as hyphens and greater-than characters (e.g. 600Val>Glu). At the DNA level, the use of hyphens along with greater-than characters was proposed (e.g. 1799T->A). When found, the rsid was also used as a synonym.

This strategy aims at retrieving a maximum of relevant abstracts. We then apply additional strategies in order to re-rank the abstracts.

2.1.2 Drug density

Our first run (*SIBTMlit1*) assumes that an abstract with a high frequency of drug names is probably more relevant to support our task, which consists to retrieve existing knowledge in the scientific literature regarding treatment of cancer. We thus use the pre-annotation of the abstracts with DrugBank to estimate the drug density of an article

(i.e. the number of occurrences of drug names in the abstract and title). Two settings have been tested for the annotations of DrugBank: the first is based on all drugs and products names and synonyms available in DrugBank, while the second is limited to drugs for cancer treatment. For this, a list of 384 DrugBank records has been defined based on different resources: cancer-related categories provided by DrugBank (e.g. *Antineoplastic agents*), the Cancer Drugs List provided by the National Cancer Institute [17], the List of Cancer Chemotherapy Drugs provided by the Navigating Care [18] and the Oral Chemotherapy Drugs List provided by CareFirst [19]. Results from the baseline run are then re-ranked based on the number of occurrences of drug names per abstract.

2.1.3 Keywords density

When manually assessing the results of the first run, we observed that our search engine was retrieving abstracts targeting the correct diseases, genes and variants. However, a consequent subset of these abstracts was not related to precision medicine studies, but other aspects, such as immunohistochemistry. Our second run (*SIBTMlit2*) assumes that the presence of some specific keywords, the so-called *positive keywords* (e.g. "treat"), in the abstract should increase the relevance of the article, while the presence of other keywords, the so-called *negative keywords* (e.g. "marker"), should decrease its relevance. Therefore, a list of positive and negative stemmed keywords was defined, based on the manual screening of a subset of retrieved articles, as well as the testing of list variations on the basic tuning set.

2.1.4 Hierarchical query expansion

Our third run (*SIBTMlit3*) assumes that an article targeting a more general (supertype) or more specific (subtype) cancer type may still be valuable from a clinical point of view. For expanding a disease to its children and parents, we used a simplified hierarchy provided by NCIt [20]. It only includes concepts in the *Neoplasm by Site* and *Neoplasm by Morphology* categories (Figure 1). Pre-annotations of the corpus with the diseases are used to retrieve abstracts concerning a parent/child disease. Results are combined with the *SIBTMlit2* run.

Liposarcoma:

Parent diseases:
- *Sarcoma*
- *Connective and Soft Tissue Neoplasm*
- *Mesenchymal Cell Neoplasm*

Child diseases:
- *Dedifferentiated Liposarcoma*
- *Mixed Liposarcoma*
- *Myxoid Liposarcoma*
- *Pleomorphic Liposarcoma*
- *Well Differentiated Liposarcoma*

Figure 1 Example of disease-based query expansion

2.1.5 Publication types

Our fourth run (*SIBTMlit4*) is based on the prioritization of the literature. Indeed, we assume that some publications, such as clinical trials, should receive higher relevance for this task. Similarly, AACR and ASCO may be considered as an important source of knowledge, as a high proportion of AACR and ASCO articles are focusing on cancer therapy and precision medicine studies. Moreover, these journals are presenting recent knowledge, sometimes not yet available in Medline abstracts. We defined four categories of publications (Table 1).

	Publication type	**MeSH term**
4	Controlled clinical trial Randomized controlled trial Pragmatic clinical trial Clinical trial Clinical trial, phase i Clinical trial, phase ii Clinical trial, phase iii Clinical trial, phase iv	Clinical Trial Clinical Trial, Phase I Clinical Trial, Phase II Clinical Trial, Phase III Clinical Trial, Phase IV Controlled Clinical Trial Randomized Controlled Trial
3	ASCO proceedings AACR proceedings	Cohort Studies Follow-Up Studies Longitudinal Studies National Longitudinal Study of Adolescent Health Prospective Studies Retrospective Studies
2	Case reports Guideline Practice guideline	Case Reports
1	Clinical study Comparative study Evaluation studies Meta-analysis Clinical conference	Clinical Study Observational Study Clinical Conference Comparative Study

Table 1 Publication types and MeSH terms used to classify articles in categories.

Publication types and MeSH terms are used to attribute an article to one category. If several categories are retrieved for a same article, the higher category is selected. If the publication belongs to one of these categories, its score is boosted. We tested different weights for each category.

2.1.6 Fusion with clinical trials task

Our last run (*SIBTMlit5*) is built on top of a run of the clinical trials task (*SIBTct5*). Our assumption is that clinical trials are an important source of information for cancer treatment. We thus decided to use the literature references available in clinical trials to re-rank our results. As no a priori tuning was possible for this task, we used the clinical trials run producing the larger number of results, thus expecting a more important impact on the results. For each clinical trial of the *SIBTct5* run, the cited PubMed articles were collected and given the same score as the citing clinical trial. For instance, the third clinical trial (i.e. NCT02571829), scored 0.782 in the *SIBTct5* run, cites five PubMed articles (i.e. 16603719, 11960696, 20601955, 23569312, 11872347), which all receives a score of 0.782. We thus obtain a list of ranked PMIDs that were then linearly combined with the *SIBTMlit4* run to produce the *SIBTMlit5* run.

2.2 Clinical trials retrieval

In this section, we describe the strategy and workflow we used for retrieving clinical trials. It consisted in four successive steps for handling four topic features.

2.2.1 Demographic features

First, we dealt with demographic features, which were easy to handle. Indeed, topic demographic features are patient's age and gender. In the clinical trial structure, there is an eligibility section, and three subsections that are gender, minimum_age, and maximum_age. Comprehensive screening of the collection showed that the gender are limited to three values (All, Male, or Female), and that ages are in a regular format (such as "18 Years", or "6 Months"). We thus designed a set of very simple rules in order to extract the age range and gender. Then, the first treatment was filtering: for each query, clinical trials that did not comply with the patient were discarded.

2.2.2 Disease mapping and filtering

The second step dealt with the disease feature. In both the topics and the clinical trials, diseases are mentioned in free text. Thus, we searched for an ontology in order to recognize and normalize diseases, and handle with synonyms (such as "colon cancer" and "colon neoplasm") or acronyms (such as "Gastrointestinal Stromal Tumor" and "GIST"). We investigated MeSH, but several diseases in topics did not find a match (such as "Pancreatic ductal adenocarcinoma" which is too specific for the MeSH). Then, we investigated the NCI thesaurus (limited to the oncology part). As all query diseases were exactly mapped, we chose to work with the NCI thesaurus.

We thus mapped NCI concepts in the clinical trials collection, thanks to simple regular expressions. This means that a term could be mapped in a longest word. For example, "sarcoma" was mapped in "liposarcoma". We first mapped diseases in the condition field, as it is where the disease is supposed to be specified. For example, for topic 1, "liposarcoma" was found in some trials in this condition field. We assume it is the most reliable. But manual screening also revealed that diseases also can appear in the keyword fields. For example, one trial has "colorectal cancer" for condition, but "colon cancer" in keywords (colon cancer is a subtype of colorectal cancer in the NCI thesaurus). We also mapped diseases in the condition_browse and mesh_term fields, which provides corresponding MeSH terms. Comments in trials advise that this MeSH mapping is done by an imperfect algorithm. We assume that keywords and MeSH terms fields provide reliable information, yet less than the condition field. Finally, we searched for diseases in the whole trial. Indeed, investigated diseases can be incorrectly mentioned in dedicated fields, but appear in the description of the study, or of the groups, or in the inclusion criteria. Yet, in these sections, some diseases also can be mentioned for a state of the art. For example, for a trial on melanoma and the KIT protein, GIST is mentioned to be possibly involved by KIT, but is not investigated by this trial. We assume that these mappings are the less reliable, but can identify relevant diseases in trials that were not retrieved by the first mappings. Yet, we discarded the exclusion criteria from these mappings.

We thus had NCI concepts mapped in the topics, and in all clinical trials (in different sections). Diseases were handled with filtering: clinical trials that did not correspond to the query condition were discarded. We decided to have more or less relaxed criteria. First, for a high precision run (SIBct1), for each topic, we only kept trials that shared the same NCI concept in the condition section. Then, we exploited the NCI hierarchy in order to include trials that shared a subtype of the topic disease in the condition field (SIBct2). Indeed, as mentioned before, a colon cancer is a subtype of a colorectal cancer. Yet, we assume that a trial dealing with a colorectal cancer is relevant for a patient with colon cancer. Then, we extended the investigated trials sections to keywords and MeSH conditions (SIBct3). Then, we also included trials that shared a supertype of the topic disease (SIBct4).

Finally, for what was supposed to be the most recall oriented run, we extended the investigated trials sections to the whole document (SIBct5).

2.2.3 Gene and variants

In pre-analysis, we first mapped gene and variant names in the clinical trials thanks to regular expressions. The output showed that the collection contained few mentions of the topic genes (on average only 271 documents per gene or variant names). Moreover, we performed manual searches in the clinicaltrial.gov engine with gene name synonyms, or variant synonyms generated by the system reported previously. But none of these synonyms were found in the collection. We assumed that this feature could be efficiently handled with Information Retrieval. Moreover, the vectorial model would favor rare words (such as gene names) that are repeated in the trial. Then, for each topic – because of the prior demographic and disease filtering – we indexed only the possibly relevant trials with Terrier, BM25 weighting scheme. Once again, the exclusion criteria were discarded, as we found examples of trials that excluded specific genes in this section.

2.2.4 Comorbidities

Finally, we used the MeSH in order to map comorbidities in the "other" topic feature, then in the exclusion criteria section of the retrieved trials. For all runs, trials that contained a comorbidity were downweighted (50% penalty) but still submitted.

Table 1 presents the five different runs that were computed for the competition.

Run	Use of NCI hierarchy for disease mapping	Fields for disease mapping	#CT per topic after filtering	#CT per topic after IR
SIBct1	No	<condition>	1362	62
SIBct2	Children included	<condition>	1801	68
SIBct3	Children included	<condition> + <conditionMeSH> + <keywords>	2040	74
SIBct4	Children and parents included	<condition> + <conditionMeSH> + <keywords>	10772	251
SIBct5	Children and parents included	all fields	26170	379

Table 1: description of the different strategies for disease filtering, and corresponding averages of clinical trials after filtering, then after Information Retrieval with gene and variants names. Runs are ranked from strictest criteria (1), which tend to favor high precision, to most relaxed criteria (5), which tend to favor recall.

3. Results & Discussion

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

3.1 Scientific abstracts retrieval

In the following, we first present the settings of the system. We then report on the results obtained in the final evaluation.

3.1.1 Tuning settings

The selection of the best settings for our system relies on the basic tuning set described in section 2.1.

The linear combination of the three different queries uses the following weights: results from Query 1 receives a weight of 65%, results from Query 2 gets a weight of 15% while results from Query 3 are attributed a weight of 20%.

Regarding the drug density run, we observed that the use of the whole DrugBank performed better than the cancer-limited list. We obtained the best results when a weight of 38% was given to the drug density parameter.

Regarding the keyword density run, the best results were obtained when using the keyword list presented in Figure 2, as well as a weight of 13% for the keyword density information.

Positive keywords:
treat; drug; therap; prognos; surviv

Negative keywords:
immuno; marker; detect; sequencing

Figure 2 Lists of positive and negative keywords

Regarding the expansion to more general and specific diseases, we obtained the best results when a weight of 3% was given to the expanded queries.

Regarding the publication types based run, the best results were obtained when only the categories 4 and 3 were used, with a boost of respectively 60% and 20% was used. However, due to the benchmark limited to PubMed articles, the impact of the category 3, which is boosting, among others, the ASCO and AACR articles cannot be really estimated.

Regarding the final run, merging results with the clinical trials task, we applied here a priori and intuitive settings. A weight of 30% was given to the PMIDs obtained through the clinical trials task.

3.1.2 Final results

Results for the 30 topics are presented in Table 2. Metrics used for this task are infNDCG, P10 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. P10 (precision at rank 10) represents the proportion of relevant documents retrieved in the top ten results. It thus reflects the ability of the system to retrieve relevant results at high ranks. Finally, R-Prec (R-Precision) return the number of relevant documents returned in the top R document, where R corresponds to the number of relevant documents for the query.

	infNDCG	P10	R-Prec
SIBTMlit1	0.400	0.520	0.257
SIBTMlit2	0.410	0.527	0.262
SIBTMlit3	0.413	0.523	0.266
SIBTMlit4	**0.418**	**0.550**	**0.269**
SIBTMlit5	0.362	0.483	0.235

Table 2: Final results for the 30 topics for the scientific abstracts task

Our first strategy resulted in an infNDCG of 0.400, a P10 of 0.520 and a R-Prec of 0.257. When using in addition the keywords density, our results are slightly improved regarding all measures, respectively of +2.3% for the infNDCG (0.410), +1.3% for the P10 (0.527) and +2.1% for the R-Prec (0.262). The third strategy, expanding diseases to parents/children diseases, has a positive impact regarding the infNDCG (+0.8%) and R-Prec (+1.6%), while the P10 slightly decreased (-0.6%). The combination with the fourth strategy (i.e. favoring articles based on the publication types) enabled to achieve our best results: the infNDCG reached 0.418 (+1.1%), the P10 increased to 0.550 (+5.1%), while the R-Prec was also improved to 0.269 (+0.9%). Finally, our last strategy, aiming at combining results produced for the clinical trial tasks with results produced for the scientific abstract task, results were not conclusive: all three measures were strongly decreased.

3.2 Clinical trials retrieval

Results for the 30 topics are presented in Table 3. During the TREC workshop, only P5, P10 and P15 metrics were displayed. If Precision at high ranks is a useful indicator, we choose to display in this Table other metrics that shows complementary aspect of the system's performances. For instance, if we assume that a user is ready to screen 200 results in order to find relevant documents, then R200 (Recall at rank 200) is interesting. R-Precision is used as a balanced metric between Precision and Recall.

	P10	R-Prec	R200
SIBct1	0.289	0.138	0.181
SIBct2	0.318	0.172	0.228
SIBct3	**0.364**	0.198	0.256
SIBct4	0.336	0.244	0.419
SIBct5	0.334	**0.259**	**0.504**

Table 3 Final results for the 30 topics for the clinical trials task

It is worth reminding that the submitted runs were supposed to be from the most strict filtering (and thus most precise) run SIBct1, to the most relaxed filtering (and thus with most recall) run SIBct5. Yet, Precision performances are not consistent with this. The most surprising is that the strictest filtering (only keeping trials where the exact disease was mapped only in the condition field) leads to the worst observed precision. Yet, Recall performances are consistent. 50% of the relevant trials can be found in the top 200 trials returned by the most relaxed strategy. Focusing in R-Prec, it seems that strategies that favor Recall show a better general behavior for this task.

Conclusion

While information regarding disease, gene and variant is usually retrieved in full text articles, scientific abstracts reporting on treatment, prognosis and prevention of cancer do not always mention all this information. Therefore, the system we developed here for the scientific abstracts task is based on a constraint relaxing strategy, aiming to retrieve a maximum number of potentially relevant abstracts. Further strategies focus on the proper ranking of the retrieved abstracts. Results showed that four out of the five strategies tested to re-rank the results were valuable regarding the evaluation metrics selected for the task. However, due to the lack of training data, the impact obtained for each strategy is probably lower than what could have been expected with a better tuning of the parameters. Indeed, the weight a-priori attributed to each re-ranking strategy was relatively low. Among our strategies, the most innovative one, which consisted in

favoring scientific abstracts cited in clinical trials, showed un-conclusive results. Further investigations regarding this strategy are planned. Indeed, the clinical trial run selected for the merging showed relatively low performance after the evaluation results were released. We can thus expect that this strategy could still be valuable if based on a better run.

For the Clinical Trials task, strategies that favor Recall show a better general behavior, but it is disconcerting that strict filtering for disease mapping has poor - or even counterproductive - impact on Precision values. It is a pity that the relevance guidelines were not published before the run submission deadline. For instance, despite informal discussions in the task's forum, it was unclear if diseases' supertypes should be considered as relevant or not, while finally they were. Beyond this, teams that performed well exploited other sections in clinical trials, such as the phase or the study type. Now that gold standard is available, it will be possible to better catch what makes a clinical trial relevant or not.

Acknowledgments

The study reported in this paper has been partially supported by the Swiss National Fund for Scientific Research, SNF Grant [neXtPresso project, SNF #153437]. This work also benefited from discussions with Daniel Stekhoven, Franziska Singer and Nora Toussaint.

References

[1] http://bitem.hesge.ch/

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

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

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

[5] J Gobeill, A Gaudinat, E Pasche, 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.

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

[7] The UniProt Consortium. "UniProt: a hub for protein information". Nucleic Acids res. 2015:43(D1):D204-12.

[8] Gaudet P, Michel PA, Zahn-Zabal M, Cusin I, Duek PD, Evalet O, Gateau A, Gleizes A, Pereira M, Teixeira D, Zhang Y, Lane L, Bairoch A. The neXtProt knowledgebase on human proteins: current status. Nucleic Acids Res. 2015 Jan;43(Database issue):D764-70.

[9] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.

[10] N Sioutos, S de Coronado, HW Haber, FW Hartel, WL Shaiu, LW Wright. "NCI Thesaurus: a semantic model integrating cancer related clinical and molecular information". J Biomed Inform. 2007:40(1):30-43.

[11] F Minguet, L Van Den Boogerd, T Salgado, C Correr, F Fernandez-Llimos. Characterization of the Medical Subject Headings thesaurus for pharmacy. American Journal of Health-System Pharmacy, 71(22), 1965-1972. 2014-

[12] V Law, C Knox, Y Djoumbou, T Jewison, AC Guo, Y Liu, A Maciejewski, D Arndt, M Wilson, V Neveu, A Tang, G Gabriel, C Ly, S Adamjee, ZT Dame, B Han, Y Zhou, DS Wishart. "DrugBank 4.0: shedding new light on drug metabolism". Nucleic Acids Res. 2014:42(1):D1091-7.

[13] MJ Landrum, JM Lee, M Benson, G Brown, C Chao, S Chitipiralla, B Gu, J Hart, D Hoffman, J Hoover, W Jang, K Katz, M Ovetsky, G Riley, A Sethi, R Tully, R Villamarin-Salomon, W Rubinstein, DR Maglott. "ClinVar: public archive of interpretations of clinically relevant variants.", Nucleic Acids Res. 2016:44(D1):D862-8.

[14] M Griffith, NC Spies, K. Kyrsiak, JF McMichael, AC Coffman, AM Danos, BJ Ainscough, CA Ramirez, DT Rieke, L Kujan, EK Barnell, AH Wagner, ZL Skidmore, A Wollam, CJ Liu, MR Jones, RL Bilski, R Lesurf, Y Feng, NM Shah, M Bonakdar, L Trani, M Matlock, A Ramu, KM Campbell, GC Spies, AP Graubert, K Gangavarapu, JE Eldred, DE Larson, JR Walker, BM Good, C Wu, AI Su, R Dienstmann, AA Margolin, D Tamborero, N Lopez-Bigas, SJM Jones, R Bose, DH Spencer, LD Wartman, RK Wilson, ER Mardis, OL Griffith. "CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer". Nat Genet. 2017:49(2): 170-174.

[15] JT Dunnen, R Dalgleish, DR Maglott, RK Hart, MS Greenblatt, J McGowan-Jordan, AF Roux, T Smith, SE Antonarakis, PEM Taschner. "HGVS recommendations for the description of sequence variants: 2016 Update". Human mutation. 2016:37(6):564-569.

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

[17] National Cancer Institute. "A to Z List of Cancer Drugs." https://www.cancer.gov/about-cancer/treatment/drugs

[18] List of Cancer Chemotherapy Drugs. NavigatingCare. https://www.navigatingcare.com/library/all/chemotherapy_drugs

[19] CareFirst. "Oral Chemotherapy Drugs." https://member.carefirst.com/carefirst-resources/pdf/oral-chemotherapy-drug-list-sum2714.pdf

[20] NCI thesaurus. "NCIt Neoplasm Core Hierarchy by Site and Morphology." "https://evs.nci.nih.gov/ftp1/NCI_Thesaurus/Neoplasm/Neoplasm_Core_Hierarchy.html