Published in Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC 2020), 16-20 November, Online, which should be cited to refer to this work.

SIB Text Mining at TREC Precision Medicine 2020

Emilie Pasche[a,b], Déborah Caucheteur[a,b], Luc Mottin[a,b,c], Anaïs Mottaz[a,b], Julien Gobeill[a,b], 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: emilie.pasche@hesge.ch

Abstract

TREC 2020 Precision Medicine Track aimed at developing specialized algorithms able to retrieve the best available evidence for a specific cancer treatment. A set of 40 topics representing cases (i.e. a disease, caused by a gene and treated by a drug) were provided. Two assessments were performed: an assessment of the relevance of the documents and an assessment of the ranking of documents regarding the strength of the evidence. Our system collected a set of up to 1000 documents per topic and re-ranked the documents based on several strategies: classification of documents as precision medicine-related, classification of documents as focused on the topic and attribution of a set of evidence-related scores to documents. Our baseline run achieved competitive results (rank #3 for infNDCG according to the official results): more than half of the documents retrieved in the top-10 were judged as relevant regarding the topic. All the tested strategies decreased the performances in the phase-1 assessment, while the evidence-related re-ranking improved performance in the phase-2 assessment.

1. Introduction

This paper describes the participation of the SIB Text Mining group [1] to the TREC 2020 Precision Medicine Track. The SIB Text Mining group, 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], TREC Deep Learning [7] and TREC Precision Medicine [8,9,10] tracks. Our group is also involved in personalized health research projects, in particular the SVIP-O (Swiss Variant Interpretation Platform for Oncology) project [11]. This project aims to harmonize variant annotations in diagnosis and to provide a centralized curated database for somatic variants from Swiss hospitals. In this project, we developed a variant-specific search engine [12] enabling triage of publications (scientific abstracts, full-texts and clinical trials). The system thus facilitates the curation of variants for personalized medicine.

The TREC PM track aims at identifying documents of interest for clinicians treating patients with cancer. While in previous years the TREC PM was based on two collections (i.e. scientific abstracts and clinical trials), only the scientific abstracts collection is used this year. The focus of the task also evolved: the objective now is to identify documents reporting on high-quality evidence. While last year the participants' systems were required to suggest treatments, the treatment is this year part of the topic. In addition, no variant mention is present in the topics.

Based on our previous TREC PM participations as well as our experience with the SVIP-O project, we adapted our system to fit with the new requirements. Our system is based on two steps: first the gathering of the documents and second the re-ranking of the documents. To collect the abstracts, we built two Elasticsearch queries: the first one strictly respects the topic (i.e. the disease, the gene and the treatment are mentioned in the retrieved documents) and the second one allows constraint relaxing (i.e. two of the three entities are mentioned in the retrieved documents). The re-ranking of the documents is initially based on strategies similar to the ones tested during previous TREC PM competitions, as well as testing of new strategies. In particular, we have investigated strategies to detect potential strong evidence in abstracts.

2. Data

2.1 Collection and topics

The collection is a snapshot of MEDLINE covering PubMed abstracts published until mid-December 2018. It corresponds to 28,137,808 citations.

The topics set is composed of 40 semi-structured synthetic cases, created with the help of precision oncologists at the University of Texas MD Anderson Cancer Center. Each topic consists of three fields: the *disease* represents the type of cancer; the *gene* contains the gene affected with a mutation and the *treatment* describes the proposed drug.

2.2 Ontologies and resources

Three publicly available ontologies have been used to normalize the topics: NCI Thesaurus for diseases, neXtProt for genes and Drugbank for treatments.

NCI Thesaurus. We used the NCI Thesaurus (NCIt) [13] 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.

neXtProt. Developed by the SIB (Swiss Institute of Bioinformatics) in 2008, the neXtProt human protein knowledgebase [14] 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.

Drugbank. The Drugbank database [15] 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.

SIBiLS. Our system relies on SIBiLS (Swiss Institute of Bioinformatics Literature Services) [16], a mirror of MEDLINE and PMC enriched with keywords and biomedical entities (about 1.4 billion) from a growing

set of standardized and legacy vocabularies. Using SIBiLS and its annotations enable first to retrieve results faster thanks to pre-computed indexes and second to increase the recall. Indeed, querying the collections through the annotations permits to retrieve not only the exact term mentioned in the topic, but also its synonyms as well as string variations. For instance, topic 4 mentions the gene BRAF. While querying MEDLINE using the keyword BRAF returns 15,907 hits, querying the SIBiLS annotations using NX_P15056 (i.e. the neXtProt unique identifier corresponding to the BRAF gene) returns 17,191 documents, thus increasing the recall by almost +8%. Indeed, the annotations pipeline recognizes not only BRAF and its official synonyms (BRAF1, RAFB1), but also syntactic variations such as B-RAF.

3. Methods

Topics are automatically pre-processed to map topic terms to unique identifiers. NCI thesaurus is used for disease mentions, neXtProt is used for gene mentions and Drugbank is used for treatments.

3.1 Baseline run

Our baseline run (sibtm_run1) is a combination of two Elasticsearch queries: an exact query and a relaxed query. The system uses the SIBiLS Elasticsearch MEDLINE index for querying. Up to 1000 documents were retrieved per query. Results are then filtered to exclude documents absent in the TREC PM collection. The exact query requires the three topic entities (i.e. the disease, the gene and the treatment) to be mentioned in the document, while the relaxed query requires two out of the three topic entities to be specified in the document. Each entity is searched through two modes: 1) the topic term is mentioned in the free text of the document or 2) the corresponding unique identifier is mentioned in the SIBiLS annotations of the document. Documents of the two queries are merged and a weight of 1.0 is accorded to Elasticsearch' scores of documents returned by the exact query, while a weight of 0.4 is attributed to Elasticsearch' scores of documents returned by the relaxed query. The weight was tuned by using the TREC PM 2019 benchmark.

3.2 Re-ranking based on precision medicine

Our second run (sibtm_run2) aims to re-rank documents based on their relevance for precision medicine: scores of documents related to precision medicine (PM) are boosted. A classifier was built using TREC PM 2018 benchmark for training and testing. This benchmark reports for each assessed document if it is PM or not PM. For some documents, different assessments were retrieved among topics. We excluded such documents. Several classifiers have been tested: k-neighbors, SVC, decision tree, random forest, gradient boosting, Ada boost, logistic regression and MLP classifiers. The best results were obtained using a k-neighbors classifier (k=4). After testing, only titles and abstracts were used for the classification and TF-IDF vectorizer was used to transform the text to vectors.

The classifier is used to attribute a binary score to documents retrieved by the baseline run (sibtm_run1) and documents are re-ranked by adding a positive boost of 0.15 to documents classified as PM. The value of the positive boost was tuned using the TREC PM 2019 benchmark.

3.3 Re-ranking based on document focus

Our third run (sibtm_run3) aims to re-rank documents based on the focus of the document regarding the topic. The TREC PM 2018 benchmark was used to build a training and testing collection to develop a classifier. The collection consists of two "focus" scores (one for the gene and one for the disease) calculated

for each topic/document pair of the TREC PM 2018 benchmark. For instance, for a gene, this score represents the percentage of occurrences of the given gene regarding occurrences of all genes in the document. Thus, a low score (e.g. the document mentions only one time the gene of interest, while one or several other genes are mentioned multiple times) will reflect that the document is not focused on the topic gene. In addition, the density of gene, drug and disease mentions are also added to the test collection. Again, several classifiers were tested and the best results were obtained using a linear SVC classifier. The classifier is used to attribute a binary score to documents retrieved by the second run (sibtm_run2) and documents are re-ranked by adding a positive boost of 0.55 to documents classified as focused on the topic. The value of the positive boost was tuned using the TREC PM 2019 benchmark.

3.4 Re-ranking based on evidence

Our fourth (sibtm_run4) and fifth (sibtm_run5) runs are based on a re-ranking of documents according to potential evidence. To this extent, we have defined a set of criteria to estimate if a document might contain a strong positive or negative evidence. The following criteria have been selected: publication type, group size, diversity of ethnic groups, diversity of genders, diversity of age groups and the presence of keywords related to the strength of the evidence. All documents retrieved by the baseline run have been processed and associated with a set of scores corresponding to each criterion. We explain below how each score was defined.

The first score corresponds to the publication type. Each publication type has been associated with a score from 0 to 4 (see Table 1), based on TREC PM guidelines as well as *a priori* knowledge. For each document, the publication type associated with the higher score was selected.

Publication type	**Score**
Retracted publication ; Video-Audio Media ; Bibliography ; Biography ; Autobiography ; Twin Study ; Address ; Retraction of Publication ; Directory	0
Journal Article ; Case Reports ; Letter ; Comment, Editorial ; Webcast ; Congress ; Historical Article ; News ; Overall ; Published Erratum ; Interview ; Clinical Conference ; Lecture ; English Abstract ; Consensus Development Conference ; Consensus Development Conference, NIH ; Corrected and Republished Article ; Introductory Journal Article ; Duplicate Publication ; Technical Report ; Portrait ; Dataset ; Legal Case ; Patient Education Handout ; Personal Narrative ; Newspaper Article ; Interactive Tutorial ; Classical Article	1
Observational Study ; Research Support, Non-U.S, Gov't ; Research Support, N.I.H., Extramural ; Research Support, U.S. Gov't, Non-P.H.S. ; Research Support, U.S. Gov't, P.H.S. ; Research Support, N.I.H., Intramural ; Research Support, American Recovery and Reinvestment Act	2
Comparative Study ; Clinical Trial, Phase II ; Review ; Clinical Trial ; Evaluation Study ; Clinical Trial, Phase I ; Validation Study ; Clinical Study ; Guideline ; Clinical Trial Protocol ; Adaptive Clinical Trial	3
Multicenter Study ; Meta-Analysis ; Systematic Review ; Clinical Trial, Phase III ; Randomized Controlled Trial ; Pragmatic Clinical Trial ; Controlled Clinical Trial ; Clinical Trial, Phase IV	4

Table 1: Association between publication types and scores

The second score refers to the number of patients (also named "group") implicated in the paper. Expressions related to the group (e.g. "n=", "population=", etc.) are searched within the abstract. Each numeric value associated with such expressions was added to a list. The sum of all numeric values found was calculated, supposing each value is a group of patients. Finally, a score was attributed to each document depending on a manually defined scale (Table 2).

Value (x)	Score
$x < 10$	1
$11 \leq x \leq 50$	2
$12 \leq x \leq 100$	3
$101 \leq x \leq 250$	4
$251 \leq x \leq 500$	5
$501 \leq x \leq 1000$	6
$x > 1000$	7

Table 2: Scale of scores depending on the x value.

As suggested in the Relevance Judgement Guideline of TREC 2020 Precision Medicine, the ethnic diversity could be of interest to judge the quality of evidence [17, 18]. The third score represents ethnic diversity in the group. The presence of ethnic groups was searched in both the free text (title and abstract) and the MeSH terms associated to MEDLINE documents. For the search in free text, a list of patterns representing several ethnic groups ("hispanic", "hispanics", "latino", "latinos", "asian", "asians", "asiatic population", "caucasian", "caucasians", "european", "europeans", "amerindian", "amerindians", "american", "americans", "african", "africans") was manually defined. Patterns corresponding to a same ethnic group are associated to a unique label (e.g. “hispanic” and “latino” will be replaced by the stemmed label “hisp”). For the MeSH search, a list MeSH concepts related to ethnicity has been selected (e.g. “D044465” for “European Continental Ancestry Group”). The score attributed to a document corresponds to the maximum distinct ethnic groups identified in the free text or in the MeSH terms. If no ethnic group term is identified in the text but the “multiracial” term is found in free text, the score is updated to 2.

The fourth score reflects gender diversity in the group. As it was done for the ethnicity, gender groups are searched in free text using a set of patterns (e.g. “male”, “female”, “woman”, ...) and in the MeSH terms using a set of MeSH concepts (e.g. “D005260” and “D008297” for “Female” and “Male” respectively). The score attributed to a document corresponds to the maximum distinct gender groups identified in the free text or in the MeSH terms.

The fifth score describes age diversity. Here also, the free text search and the MeSH terms (e.g. “D000293” for “Adolescent” or “D000328” for “Adult”) search were performed to identify the age groups. For the free text, two scores were allocated to this strategy, a first one based on the matching of numeric values (e.g. “age=xxx”) normalized in subsets (Table 3), another one based on the matching of “categories” (e.g. “child”, “adult”) also normalized in subsets. The maximum number of different subsets found was selected to represent the age diversity.

Age (x)	Normalized in...
$x < 2$	baby
$2 \leq x \leq 12$	child
$13 \leq x \leq 18$	teenager
$19 \leq x \leq 44$	adult
$45 \leq x \leq 64$	middle aged
$65 \leq x \leq 79$	aged
$x \geq 80$	senior

Table 3: Normalization of age for the numeric matching strategy.

Finally, the last score is based on keywords manually judged as related to strong evidence (Table 4). For each of them, a score was associated and summed to obtain the relevance score of each paper.

Keyword	Score
certain ; uncertain	1
success ; fail ; strong ; weak ; validation	2
in vivo	3
relevant ; irrelevant ; pvalue ; significant ; significative	4

Table 4: Association between pertinent keywords and scores.

The fourth run (sibtm_run4) is based solely on the publication type score and the strength of evidence score, which are in our opinion the strongest parameters to assess evidence. Both scores are merged together to generate the evidence_score. Documents are re-ranked by adding a 0.2*evidence_score to each document returned by the third run (sibtm_run3).

The fifth run (sibtm_run5) is based on all criteria scores. Each score receives an *a priori* weight (i.e. 0.6 for publication types, 0.4 for strength of evidence, 0.4 for size groups, 0.2 for diversity of ethnic groups, 0.4 for diversity of gender groups and 0.4 for diversity of age groups) to generate the evidence_score. Documents are re-ranked by adding a 0.2*evidence_score to each document returned by the third run (sibtm_run3).

4. Results and discussion

The assessment of the TREC PM results was conducted in two phases: 1) ranking assessment and 2) evidence assessment. In phase-1, assessors evaluated whether the document was relevant given the topic. A document was judged as relevant if it was about precision medicine and if the topic entities (i.e. the disease, the gene and the treatment) were discussed in the document. In phase-2, assessors evaluated whether the document reported strong evidence or not. To this extent, 5 levels of evidence were defined for each topic: 0 for no evidence in the paper and 4 for best-available evidence. The assessors then assigned the documents to one of these levels. A gain value was then computed for each document: 0 if the level is 0 and 2^(level-1) otherwise.

Metrics used to evaluate the first phase 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. Metric used to evaluate the second phase is NDCG@30. It represents the gain brought by the first 30 documents based on their position in the ranked results.

Results for the 40 topics are presented in Table 5. The baseline run achieves competitive results: it obtains an infNDCG of 52.76%, which ranks it #3 out of 16 participants. However, surprisingly, all the further strategies strongly decrease the performance. The strategy to detect precision medicine-related documents decreases the infNDCG by -10%, while the strategy to favor documents discussing mainly the entities of the topics decreases the infNDCG by -12.4%. The two strategies to re-rank documents based on evidence have also a negative impact, however less strong: of -1.9% for the run 4 and -5.3% for the run 5. We could

explain these results by the fact that the tasks of TREC PM 2018 and TREC PM 2019, which benchmarks were used to tune the systems, strongly differ from this year' tasks. Especially for run 2 and run 3, the large number of wrong predictions is resulting from the supervised learning not capable of generalizing on new data.

Results of the phase-2 assessment however shows that our two strategies to favor documents with best-quality evidence has a positive impact on the results. The strategy of run 4, which is using two criteria to define evidence, enables improving the NDCG@30 by +6.2%. The second strategy based on six criteria has a stronger positive impact by improving NDCG@30 by +13.6%. However, the evidence strategies were tested on top of our weaker run, thus resulting in final performances below the median participants' performances. We could expect that such strategies applied on top of our best run (i.e. the baseline run) would result in more competitive results.

	Phase-1 assessment			**Phase-2 assessment**
	infNDCG	**P@10**	**R-Prec**	**NDCG@30**
sibtm_run1	0.5276	0.5323	0.4020	0.2756
sibtm_run2	0.4747	0.4742	0.3703	0.2239
sibtm_run3	0.4156	0.4484	0.2907	0.1736
sibtm_run4	0.4077	0.4452	0.2774	0.1845
sibtm_run5	0.3860	0.4355	0.2631	0.1972

Table 5: Resume of the results for the five submitted runs for phase-1 and phase-2 assessments.

5. Conclusion

At this final edition of TREC PM, our baseline system achieved competitive results: more than half of the results returned in the top-10 are judged relevant to the topic. However, all our "advanced" strategies resulted in a strong decrease of the performance in the abstract triage assessment. We assume that this loss is a consequence of our tuning relying on the benchmarks from the previous TREC PM sessions, which are eventually too dissimilar from the topics/objectives of 2020.

Nevertheless, the re-ranking strategies focusing on evidence showed a positive impact to detect documents with high quality evidence. Our strategy was based on manual screening of the literature to define criteria. All settings were *a priori* defined. We can now envisage to improve this approach by using the TREC PM 2020 relevance judgment to automatically tune the best parameters.

Acknowledgments

This experiment has been supported by the Swiss Personalized Health Network (SPHN) and BioMedIT fundings (see https://svip.ch/). This work also benefited from discussions with the SVIP members.

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] J Knafou, M Jeffryes, L Mottin, D Teodoro and P Ruch. SIB Text Mining at TREC 2019 Deep Learning Track: Working Note. In TREC. 2019.
[8] 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.
[9] 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.
[10] D Caucheteur, E Pasche, J Gobeill, A Mottaz, L Mottin and P Ruch. Designing retrieval models to contrast precision-driven ad hoc search vs. recall-driven treatment extraction in Precision Medicine.
[11] DJ Stekhoven, P Ruch and V Barbié. Swiss Variant Interpretation Platform for Oncology (SVIP-O), Swiss Med Informatics 34 (2018), 00411.
[12] 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
[13] 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.
[14] 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.
[15] 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.
[16] J Gobeill, D Caucheteur, PA Michel, L Mottin, E Pasche and P Ruch. SIB Literature Services : RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts. Nucleic Acids Res 48(W1) (2020), W12-W16.
[17] C Grenade, MA Phelps, MA Villalona-Calero. Race and ethnicity in cancer therapy: what have we learned? Clinical pharmacology and therapeutics (2014), 95(4), 403-412
[18] LT Clark, L Watkins, IL Piña, M Elmer, O Akinboboye, M Gorham, B Jamerson, C McCullough, C Pierre, AB Polis, G Puckrein, JM Regnante. Increasing Diversity in Clinical Trials: Overcoming Critical Barriers (2019), *44*(5), 148–172.