SIB Text Mining at TREC Clinical Trials 2021

Déborah Caucheteur[a,b], Emilie Pasche[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: deborah.caucheteur@hesge.ch

Abstract

TREC 2021 Clinical Trials Track aimed to develop algorithms to improve patient recruitment in clinical trials. These recruitment problems represent a real obstacle to medical research, leading to delays in clinical trial schedules and sometimes even to the termination of the trial due to the lack of eligible patients recruited. A set of 75 topics was distributed to participants. Each topic represents a patient's medical record, in other words a patient case description in free text format. In parallel, a set of clinical trials from ClinicalTrials.gov was also provided. The challenge was then to determine for each patient, if during a recruitment phase for a clinical trial from the corpus, the patient would be assessed as eligible, excluded or irrelevant.

As an output, for each topic, our system returns a list of clinical trials ranked from the highest (relevant) score to the lowest within the limit of 1,000 results per topic. In total, five strategies were tested corresponding to the five runs submitted and will be described in this paper.

The publication of the results at the TREC conference in November 2021 will indicate whether one of the strategies has proved more likely to deliver good results or whether, on the contrary, the work should be redirected towards new ideas.

1. 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], TREC Deep Learning [7, 8] and TREC Precision Medicine [9, 10, 11, 12] tracks. This paper describes our participation in the TREC 2021 Clinical Trials track.

Among the projects in which our group is involved, we can mention the SVIP-O (Swiss Variant Interpretation Platform for Oncology) project [13] which 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 [14] enabling triage of publications (scientific abstracts, full-texts and clinical trials). The system thus facilitates the curation of variants for personalized medicine. So we therefore have experience with scientific publications, including clinical trials, as well as ontologies useful in the biomedical field [15], and a robust infrastructure called SIBiLS for SIB Literature Services [16] with specific pipelines to try to address the issue.

The TREC Clinical trials track aims to identify clinical trials for which patients (described in a topic) would be considered relevant in the study recruitment process. Thanks to our previous TREC participations, especially in the Precision Medicine track, we used the same approach based - firstly on the gathering of the documents: the clinical trial archives were imported from TREC website and loaded into our MongoDB, the database software we use repeatedly in our pipelines, and the list of topics (patients); - secondly on the re-ranking of the documents with a Lucene index engine.

The topics mostly describe the patient's age, gender, lifestyle, serology results, etc... Unlike previous Precision Medicine tracks, this year's topics did not contain any gene (except for Topic 68) or variant names. We therefore decided not to repeat the annotation process specific to genes, as it was of no interest, but to keep it for the terminologies relating to diseases in particular.

To retrieve the clinical trials, we used Elasticsearch queries including some variations depending on the adopted strategy for each run.

2. Data

2.1 Collection and topics

The collection is an archive of 375,580 clinical trials in XML format, published on ClinicalTrials.gov between November 1999 and April 2021, available for importation on TREC website.
The topics list contained 75 patient descriptions with several data, based on the use of EHRs (Electronic Health Record) whose storage is useful for routine medical care.

2.2 Ontologies and resources

To normalize the topics and construct the annotations of the clinical trials, four publicly available ontologies have been used: ICD-10, SNOMED CT, MESH and NCI Thesaurus.

NCI Thesaurus. Provided by the National Cancer Institute, the NCI Thesaurus (NCIt) [17, 18] is a standard for biomedical coding and reference, used both by public and private scientific partners worldwide.We used this terminology for disease mapping. It covers several fields like translational and basic research, clinical care, public information and administrative activities.

SNOMED CT. SNOMED Clinical Terms [19, 20] or SNOMED CT, is a computer processable collection of medical terms maintained and distributed by SNOMED International, an international non-profit standards development organization. It provides codes, terms, synonyms and definitions used in clinical documentation and reporting. As the terminology is multilingual, only the English language, that of the clinical trials collection, has been retained.

MESH. Provided by the U.S. National Library of Medicine (NLM), the Medical Subject Headings (MeSH) [21, 22] 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.

ICD-10. Provided by the World Health Organization (WHO), ICD-10 [23, 24, 25] is a medical classification, the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD). This classification contains codes for diseases,

signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases.

To store and search in the collection, MongoDB and Elasticsearch have been used.

MongoDB. As a non-relational database management system, MongoDB [26] uses flexible documents instead of tables and rows to process and store various forms of data. It allows queries and analyses of large amounts of information. MongoDB documents are organised into collections, formatted as Binary JSON and can be distributed across multiple systems.

ElasticSearch. Built on Apache Lucene and developed in Java, ElasticSearch [27] is an open-source search and analytics engine that allows to store, search and quickly analyze huge volumes of data because it searches an index instead of searching in the full text.

3. Strategies

3.1 Topics and collection pre-processing

A pre-processing of the topics and the clinical trials collection is performed before query time.

Clinical trials collection

For each clinical trial, some fields of interest are selected and loaded into a MongoDB collection: the titles (brief and official), brief summary, detailed description, condition, intervention (type, name, description), criteria (including inclusion and exclusion), gender, age (minimum and maximum, with a conversion in “days”), mesh terms, keywords...

Then, these fields are annotated in order to attribute codes from various terminologies for quicker queries, i.e. NCI Thesaurus, ICD-10, MESH and SNOMED. This process implies string pre-processing and tokenization methods, both applied on the clinical trials and on the terminologies. A dash or a slash could sometimes be responsible for non-matching. In our pipeline, this risk is removed thanks to the creation of 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 occurences of the word with the dash are present. The annotations are then pushed into a new MongoDB collection, with one document per clinical trial.

Finally, each document (clinical trial) is indexed in an ElasticSearch index (version 7.13.4).

Topics

In parallel, topics are also pre-processed to map terms to unique identifiers based on terminologies previously described. The fields are extracted and restructured in a JSON format as:

- “topic_number”,
- “age”: thanks to regex and arithmetic operations, normalisation in days to distinct children and adults;
- “gender”: normalisation in “male” or “female”;
- “free_text”: keep the original description;
- “diseases_annotations”: merge the annotations from SNOMED CT, ICD-10 and NCI Thesaurus terminologies;
- “snomed_annotations”: specific annotations constructed on SNOMED CT terms;
- “icd_annotations”: specific annotations constructed on ICD-10 terms;
- “nci_annotations”: specific annotations constructed on NCI Thesaurus terms;
- “mesh_annotations”: specific annotations constructed on MESH terms.

The aim to keep separate annotations from different terminologies is to be able to filter on one specific origin later if needed.

The output is a JSON file available for the next query step, the clinical trials retrieval.

3.2 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.

Five queries, one per run, are designed as presented in Table 1. Differences between these queries are about constraints relaxing and boosting on specific fields or codes. Settings have been arbitrarily defined.

All queries share a common “basis”:

- the clinical trial must address the patient’s gender or address all genders;
- the patient’s age must be included between the minimal and the maximal age of admission in the clinical trial;
- The topic’ annotation codes must not be found into the exclusion criteria field of the clinical trial. If at least one code is found, the clinical trial is automatically rejected from the list.

Query 1 (SIBTMct1) is the most restrictive request based on the presence of at least one ICD code shared between the topic and the clinical trial. NCI Thesaurus, SNOMED and MESH codes are not mandatory. Boosts are different depending on the code sources: NCI code > MESH code > SNOMED code. Also, a major boost is applied if codes are found in the inclusion criteria field.

Queries 2, 3, 4 and 5 (SIBTMct2/3/4/5) are less constrained than the first query. Indeed, the presence of codes from a specific source is no longer mandatory (previously ICD-10 was required). Differences between these queries are based on the boosting: depending on the code sources, ICD code > NCI code > MESH code ; and a major boost applied if codes are found in the inclusion criteria field + keywords / MESH / description detailed for query 3 / 4 / 5 respectively.

		Query 1	Query 2	Query 3	Query 4	Query 5
Common basis						
mandatory conditions	Gender	exact or “all”				
	Min. age	Inferior or equal to patient’s age				
	Max. age	Superior or equal to patient’s age				
	Exclusion criteria	Not include an annotation code shared with the topic				
Query details						
Presence of codes	ICD	must	should			
	NCI	should				
	SNOMED	should				
	MESH	should				
Boost on codes	ICD	-	50			
	NCI	20	30			
	MESH	10				
Boost on fields	Inclusion criteria	100				
	Keywords	1	1	50	1	1
	MESH	1	1	1	50	1
	Detailed description	1	1	1	1	50

Table 1: Resume of the queries - common basis and details about presence requisition and boosts on codes or fields.

4. Results and Discussion

For his first edition, the TREC Clinical Trials 2021 track received a total of 113 runs from 26 different teams; 12 of the runs were manual runs and 101 were automatic runs.

During the evaluation phase, for each of the 75 topics, each document (clinical trial) received a judgement among 0, 1 or 2:

- 0 as “not relevant”;
- 1 as “excluded”: the patient is not eligible for the trial due to the exclusion criteria but the trial is indeed related to the patient’s condition;
- 2 as “eligible”: the patient described in the topic is eligible for this clinical trial.

For measures based on binary judgments, only eligible is treated as relevant. The NDCG measure is computed using gains of 1 for excluded and 2 for eligible.

Metrics used to evaluate the document ranking are presented below and the averages of these metrics for each run (SIBTMct1 to SIBTMct5) are shown in Table 2.

- NDCG@10: it represents the gain brought by the first 10 documents based on their position in the ranked results.
- Prec@10: Precision at 10 is the proportion of the top-10 documents that are relevant [28]. It thus reflects the ability of the system to retrieve relevant results at high ranks.
- Reciprocal Rank: The Reciprocal Rank (RR) information retrieval measure calculates the reciprocal of the rank at which the first relevant document was retrieved. RR is 1 if a relevant document was retrieved at rank 1, if not it is 0.5 if a relevant document was retrieved at rank 2 and so on [29].

	Measures		
Run	**NDCG@10**	**Prec@10**	**Reciprocal Rank**
SIBTMct1	0.249	0.167	0.458
SIBTMct2	0.255	0.176	**0.469**
SIBTMct3	0.268	0.192	0.453
SIBTMct4	0.258	0.184	0.451
SIBTMct5	**0.278**	**0.199**	0.452

Table 2: Resume of the results for the five submitted runs.

The baseline run (SIBTMct1) which is the most restrictive displays the worst metrics. The hypothesis to relax constraints and apply boosts on some codes and fields applied to the next runs improve the clinical trials retrieval.

The SIBTMct2 obtains the best RR value (0.469) and the SIBTMct5 obtains the best NDCG@10 and Prec@10 values (respectively 0.278 and 0.199) with a RR value not so much lower (0.452 vs 0.469).

Before the TREC conference, a summary table shows per-topic min/max/median values for the three measures (NDCG@10, Prec@10, Reciprocal Rank). With these datas, for each topic, the measures obtained in each run were compared with the relative median and best value among all the participant's runs, then counted to obtain a value on the 75 topics as presented in Table 3.

SIBTMct5 seems to be the best with 52% of NDCG@10 values, 74.4% of Prec@10 and 57.3% of RR values equal or superior to the respective median values, and 25/75 topics for which we obtained the best RR value (equivalent to 33.3%). It could be discussed with SIBTMct2, where Reciprocal Rank has the highest percentage (60% superior or equal to the median, 28/75 topics with the best value), a Prec@10 not much inferior to the highest (73.3%) but a NDCG@10 value equals to 42.7.

	First analysis of our results					
	NDCG@10		**Prec@10**		**Reciprocal Rank**	
Run	≥ median	= best	≥ median	= best	≥ median	= best
SIBTMct1	38.7	0	66.7	1.3	57.3	36
SIBTMct2	42.7	0	73.3	0	**60**	**37.3**
SIBTMct3	50.7	0	**74.7**	1.3	58.7	33.3
SIBTMct4	44	0	73.3	0	**60**	34.7
SIBTMct5	**52**	0	**74.7**	1.3	57.3	33.3

Table 3: First comparison between our results and those obtained by other participants.

The major relaxing constraints between the baseline and the four other strategies seems to induce better scores. Then, by playing with boosts on particular fields or codes, results change and will be explored in more detail after the TREC conference.

5. Conclusion

The results presented at the conference showed that our runs did not place us in the Top-10 this year [30]. A study of the results should enable us to determine why and how we could have improved our performance.

Several approaches will be explored:

- Are the retrieved papers good?
- Was the score of our runs affected by papers that should not have been retrieved or was it their ranking that was not good?
- Was our approach to excluding clinical trials based on the "exclusion criteria" field too drastic¨?
- Should we have used one or more other terminologies for our annotations to capture more terms from the topics?
- For those topics for which we have not retrieved any relevant papers, what is the cause? How can this problem be solved?

About this last point, our first study highlighted a number of topics for which NDCG@10 and P@10 metrics and those of the participating teams remained mostly at 0, for example topics 9, 11 and 44. Determining how to improve our metrics on those topics that appear more complex may allow us to improve the scores of less atypical topics as well.

For the TREC Clinical Trials track edition 2022, there are no major changes planned, although the collection of clinical trials will be expanded. We can use the studies of the results of this TREC 2021 to improve our research performance.

6. 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 Jeffreyes, L Mottin, D Teodoro, and P Ruch.SIB Text Mining at TREC 2019 Deep Learning Track: Working Note. In TREC 2019.
[8] J Knafou, M Jeffreyes, S Ferdowsi, and P Ruch.SIB Text Mining at TREC 2020 Deep Learning Track. In TREC 2020.
[9] 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.
[10] 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.
[11] 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. In TREC 2019.
[12] E Pasche, D Caucheteur, L Mottin, A Mottaz, J Gobeill, and P Ruch. SIB Text Mining at TREC Precision Medicine 2020.In TREC 2020.
[13] DJ Stekhoven, P Ruch and V Barbié. Swiss Variant Interpretation Platform for Oncology (SVIP-O), Swiss Med Informatics 34 (2018), 00411.
[14] E Pasche, A Mottaz, D Caucheteur, J Gobeill, P-A Michel, and P Ruch. Variomes: a high recall search engine to support the curation of genomic variants. bioRxiv (2021).
[15] 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.
[16] J Gobeill, D Caucheteur, P-A 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 Research, Volume 48, Issue W1, 02 July 2020, Pages W12–W16, https://doi.org/10.1093/nar/gkaa328
[17] 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.
[18] https://ncithesaurus.nci.nih.gov/ncitbrowser/
[19] https://www.snomed.org
[20] M Q Stearns, C Price, K A Spackman, and A Y Wang. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp 2001: 662-666.
[21] 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.
[22] https://meshb.nlm.nih.gov/search
[23] World Health Organization. History of the development of the ICD. http://www.who.int/classifications/icd/en/HistoryOfICD.pdf. Accessed March 24, 2020.
[24] World Health Organization . *International Statistical Classification of Diseases and Related Health Problems: Instruction Manual.* Vol. 2. Geneva: WHO; 2011. 10th revision.
[25] https://www.who.int/standards/classifications/classification-of-diseases
[26] https://www.mongodb.com
[27] https://www.elastic.co
[28] Craswell N. (2009) Precision at n. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_484
[29] Craswell N. (2009) Mean Reciprocal Rank. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_488
[30] Roberts K, Demner-Fushman D, Voorhees E M, Bedrick S, Hersh W, R (2021) Overview of the TREC 2021 Clinical Trials Track. https://trec.nist.gov/act_part/conference/papers/Overview-CT.pdf