DOSSIER at TREC 2021 Clinical Trials Track

Wojciech Kusa[1] and Yasin Ghafourian[1,2]

[1] TU Wien, Vienna, Austria
wojciech.kusa@tuwien.ac.at
[2] Research Studios Austria, Vienna, Austria
yasin.ghafourian@researchstudio.at

Abstract. This paper describes our experimental setup and results for the Clinical Trials Track at TREC 2021. In particular, we study (i) the effectiveness of post-processing with patients' metadata and (ii) the novel re-ranking formula using eligibility criteria. We find that the post-processing improves the model over the baseline run. However, the custom re-ranking negatively impacts average results.

1 Introduction

This paper presents an overview of the DoSSIER team's submissions to the TREC 2021 Clinical Trials Track[1]. The main objective of the system was to retrieve all relevant clinical trials to the given 75 topics, which represent patients' descriptions. The clinical trials dataset consisted of a snapshot of a ClinicalTrials.gov website, with around 370 thousand different trials. Clinical trials are research studies designed to assess medical, surgical, or behavioural interventions. They contain lengthy descriptions and inclusion/exclusion criteria which are utilized to evaluate the trial-patient eligibility [1]. On the other hand, patients' descriptions contain a free text case description of a patient, a simulation of an admission statement in an Electronic Health Record (EHR) format.

We submit three runs coming from subsequent stages of the same architecture. We study the effectiveness of post-processing with patients' metadata compared to using baseline BM25 model. We try to model the rule that the most relevant trials should be that for which the patient's history overlaps with all inclusion criteria and, at the same time, with none of the exclusion criteria. We use a custom re-ranking formula that utilises inclusion and exclusion criteria to score the trials.

2 Methodology

Figure 1 describes the overall architecture and workflow of the system we used for all three submission runs.

Clinical trials are semi-structured documents containing multiple sections, e.g., title, summary, detailed description, or eligibility criteria. Both trials and

[1] https://www.trec-cds.org/2021.html

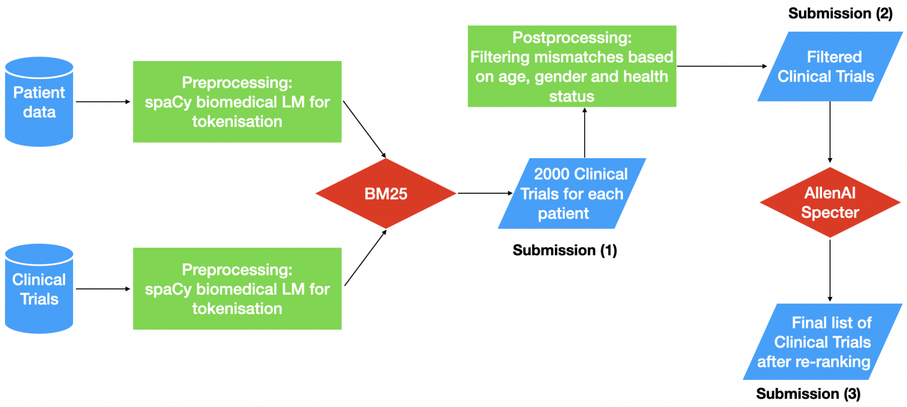

Patient data
Clinical Trials
Preprocessing: spaCy biomedical LM for tokenisation
Preprocessing: spaCy biomedical LM for tokenisation
BM25
2000 Clinical Trials for each patient
Submission (1)
Postprocessing: Filtering mismatches based on age, gender and health status
Submission (2)
Filtered Clinical Trials
AllenAI Specter
Final list of Clinical Trials after re-ranking
Submission (3)

Fig. 1. General workflow of the system we used for all three submission runs.

topics are written in complex language and contain medical jargon. We preprocess documents and topics using the ScispaCy *en_core_sci_sm*[2] biomedical language model [2]. We concatenate title, brief summary and eligibility criteria fields into a single text that represents every clinical trial. We index that data using the BM25Okapi model from the *rank_bm25* Python package[3]. After indexing, 2,000 best clinical trials for each patient that best match them are retrieved.

We manually annotate topics to extract age, gender and health status information from each patient description. This information is available in the description of the clinical trials in separate fields, which requires only parsing the XML file. We normalise these three concepts: age is a floating-point variable, and both gender and health status are binary categories.

Inclusion and exclusion criteria are mentioned in a semi-structured or unstructured format inside the trial's eligibility section. We extract this information using a rule-based approach. We successfully extracted both inclusion and exclusion criteria for 91% of documents. We split the inclusion/exclusion texts into a list of concepts. When the extraction of inclusion criteria is not possible, we utilise the full text of the trial as inclusion information. If we cannot extract exclusion criteria, we assume an empty text.

2.1 baseline

The first submission consists of 1,000 top-ranked clinical trials returned by the BM25 model for each topic. We use the default parameters for the BM25 algorithm.

[2] https://allenai.github.io/scispacy
[3] https://pypi.org/project/rank-bm25/

2.2 postproc

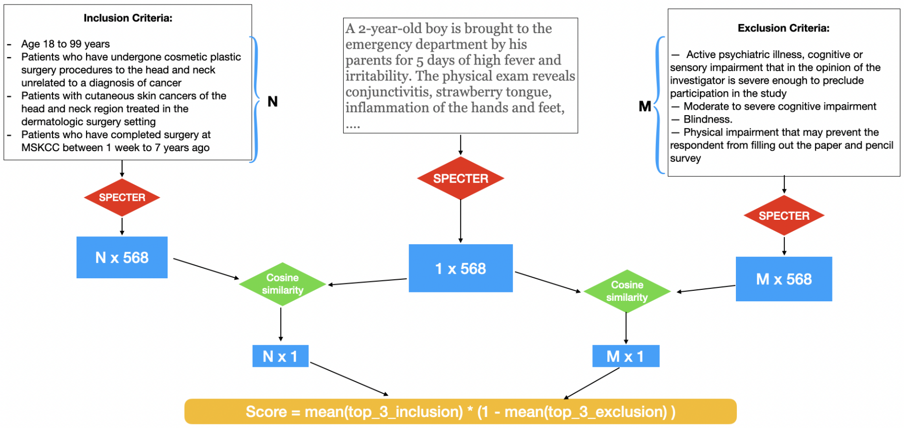
The top 2,000 trials retrieved by the BM25 model are post-processed using age, gender and health status data extracted from patients descriptions. Retrieved trials are filtered based on the stored metadata to remove the mismatches. Using the filtering, we remove on average 34% of potentially ineligible trials.

2.3 rerank

Our third submission uses as an initial pool the output from the *postproc* run. It utilises a custom formula for re-ranking using SPECTER language model [3]. An overview of the re-ranking formula is presented in Figure 2. In the first step, we take the lists with inclusion and exclusion criteria and encode them one by one using SPECTER. We do the same for the patient description. We calculate the cosine similarity between inclusion criteria and the patient encoding and also exclusion criteria and patient. We sort both of them and take the three highest inclusion and exclusion similarity scores. We calculate the final score using the following formula:

$$score = \frac{\sum_{i=1}^{3} IN_i}{3} \cdot \left(1 - \frac{\sum_{i=1}^{3} EX_i}{3}\right)$$

where IN_i means i-th most similar inclusion criterion score and EX_i means i-th most similar exclusion criterion. This is based on the reasoning that the patient's description should overlap with inclusion criteria and be as much distinct as possible to the exclusion criteria.



Inclusion Criteria:

- Age 18 to 99 years
- Patients who have undergone cosmetic plastic surgery procedures to the head and neck unrelated to a diagnosis of cancer
- Patients with cutaneous skin cancers of the head and neck region treated in the dermatologic surgery setting
- Patients who have completed surgery at MSKCC between 1 week to 7 years ago

A 2-year-old boy is brought to the emergency department by his parents for 5 days of high fever and irritability. The physical exam reveals conjunctivitis, strawberry tongue, inflammation of the hands and feet, ...

Exclusion Criteria:

— Active psychiatric illness, cognitive or sensory impairment that in the opinion of the investigator is severe enough to preclude participation in the study
— Moderate to severe cognitive impairment
— Blindness.
— Physical impairment that may prevent the respondent from filling out the paper and pencil survey

Fig. 2. Description of the neural re-ranking process.

3 Results

Results are presented in Table 1. Even though the extraction was conducted only with heuristics, post-processing with patient metadata improves all scores when compared to the baseline submission. Our custom re-ranking not only does not improve on the previous run but even further lowers the scores for all metrics when compared to the baseline. We assume that the choice of the formula to calculate the final re-ranking score had the most significant negative impact on the last submission. Due to the time and resource constraints, we could not test other formula variations.

Table 1. Performance comparison for submitted runs with across query averaged median values.

Metric	TREC's Median	baseline (1)	postproc (2)	rerank (3)
NDCG@5	—	0.401	**0.455**	0.322
NDCG@10	0.304	0.380	**0.413**	0.293
PREC@10	0.161	0.208	**0.252**	0.187
Reciprocal Rank	0.294	0.406	**0.478**	0.368

To check to what extent each step improves on the previous ones, we tested the pairwise difference between NDCG@10 scores for each topic (Table 2). The second run (*postproc*) achieves higher NDCG@10 scores for 44 and 52 topics than runs 1 and 3. Surprisingly, *baseline* obtained higher NDCG@10 over *postproc* for 13 topics. We believe that these might be caused by inaccurate extraction of the metadata from topics and clinical trials. Even though *rerank* achieved the worse mean scores, it still improved on 21 topics compared to *postproc*. Further analysis would be needed to check if this improvement is a meaningful gain from the re-ranking retrieving new relevant documents.

Table 2. Aggregation of pairwise inter-run comparisons for single topics for NDCGG@10 metric.

Run A : Run B	Run A higher	Equal scores	Run B higher
baseline (1) : postproc (2)	13	18	**44**
baseline (1) : rerank (3)	**45**	3	27
postproc (2) : rerank (3)	**52**	2	21

4 Conclusion

We submitted three runs to the Clinical Trials track of TREC 2021 and studied the effectiveness of post-processing with patient's metadata and a novel neural re-ranking formula. We found that the post-processing improves the model performance over the baseline run, but the custom re-ranking negatively impacts average results. In a future iteration of this system, we plan to evaluate other re-ranking formulas using the eligibility criteria.

Acknowledgements. This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721).

References

1. Embi, P. J., Jain, A., and Harris, C. M. (2008). Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. BMC medical informatics and decision making, 8(1), 1-8.
2. Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019, August). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In Proceedings of the 18th BioNLP Workshop and Shared Task (pp. 319-327).
3. Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. (2020, July). SPECTER: Document-level Representation Learning using Citation-informed Transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2270-2282).