UNIMIB at TREC 2022 Clinical Trials Track

Georgios Peikos and Gabriella Pasi

Information and Knowledge Representation, Retrieval, and Reasoning (IKR3) Lab,
Department of Informatics, Systems, and Communication (DISCo),
University of Milano-Bicocca, Milan, Italy
{georgios.peikos, gabriella.pasi}@unimib.it

Abstract. This notebook summarizes our participation as the UNIMIB team in the TREC 2022 Clinical Trials Track. In this work, we extend our last year's participation by further investigating the retrieval performance achieved by our decision-theoretic model for relevance estimation. Specifically, our objective was to investigate the effectiveness of our decision-theoretic model by heavily penalizing those clinical trials for which the patient has high topical similarity to the exclusion criteria. The model has been employed to estimate relevance in two retrieval settings, ranking and re-ranking. The obtained results showed that the proposed model performs equally well in both of them, while the best results in terms of precision were achieved in the re-ranking setting.

1 Introduction

The TREC 2022 Clinical Trials track is a continuation of the 2021 Clinical Trials track, which uses the same document collection and follows the same search task definition. Specifically, given a patient's synthetic case in the form of an admission note, the system's goal is to retrieve eligible clinical trials, i.e., those for which the patient meets the inclusion criteria and not the exclusion criteria. However, a trial's eligibility criteria (inclusion and exclusion) are semi-structured, while the provided synthetic cases (queries) are unstructured. This retrieval task is complex since the topical relevance to a document's part, the one that mentions the exclusion criteria, contributes negatively to the overall document's relevance.

This report outlines our team's (UNIMIB) submission to TREC 2022 Clinical Trials Track, a continuation of our last year's participation [1]. It presents the experimental setup and compares our results to the reported TREC's median performance. This year, we investigated the retrieval performance achieved by our decision-theoretic model introduced in [1].

2 Methodology

This section describes our methodology that comprises three steps, including (i) an information extraction step, in which valuable information present in a clinical trial document is extracted and indexed; (ii) a step that involves the estimation of three topical relevance scores that are associated with each clinical trial; and (iii) a score aggregation step, from which the final document ranking is obtained.

2.1 Information Extraction & Indexing

Clinical trials are structured documents with various fields such as title, summary, studied condition, among others. In addition, in this task, a trial's inclusion and exclusion criteria are mentioned in a semi-structured format within its eligibility section, holding great importance.

Our methodology exploits four document representations, each containing different document information. In detail, using a set of regex rules that leverage the semi-structured format of the inclusion and exclusion criteria, we extract them to create two distinct indices. Specifically, the *I_in* index contains only a trial's inclusion criteria, and the *I_ex* index only the trial's exclusion criteria. In the cases where their extraction was not feasible, the whole eligibility section has been

indexed in both indices, i.e., the *I_in* and *I_ex*. In addition to those two indices, we construct a third one. Here, the title, description, studied condition, and summary sections were combined in a single text and indexed; this index will be referred to as *I_main*. Lastly, we indexed all document sections to create the *I_comb* representation used in one of our experiments.

2.2 Relevance Estimation & Score Aggregation

During retrieval, given a query, we estimate for each document three topical relevance scores using the *I_in*, *I_ex*, and *I_main* indices. These scores represent the degree to which a patient's information is met in a document's inclusion and exclusion criteria, and in its main fields.

Then, the final stage of our methodology involves aggregating the obtained topical relevance scores by explicitly considering their independent contribution to relevance. Specifically, we employed TOPSIS [2], a multi-criteria decision-making method, to obtain a final document ranking. This method associates three weights with the distinct topical relevance scores obtained in the previous stage. These weights indicate how significant the similarity to the associated document's part is for the document's overall relevance. In addition, this method associates to each distinct topical relevance score an objective. This objective is either positive or negative; a positive objective indicates that this score impacts positively a document's overall relevance, i.e., the higher the score the better, while a negative objective is the exact opposite. In this work, the topical relevance score obtained from the similarity to a trials main parts and inclusion criteria are assigned a positive objective, while the topical relevance score obtained from the similarity to a trials exclusion criteria a negative objective. Therefore, due to the associated objectives, the employed aggregation method always penalizes a clinical trial for which the topical relevance score associated with its exclusion criteria is high. As a result, documents that have high topical relevance scores to a trial's exclusion criteria will obtain a lower ranking position. Considering an example, setting the importance weight of the inclusion criteria equal to .2, the exclusion criteria equal to .6 and the main parts equal to .2 leads to a retrieval system that heavily penalizes those trials for which a patient's information has high similarity to their exclusion criteria.

3 Experiments

We have submitted five runs to investigate the impact of importance weights in the relevance estimation. To index the collection and estimate the topical relevance scores for each document representation, we have employed PyTerrier [3] and the ln_expB2 divergence from randomness model [4]. For preprocessing, we have used the standard PyTerrier pipeline, i.e., porter-stemming and stopwords removal.

The first run, namely *IKR3_BSL* exploits the ln_expB2 model and the *I_comb* index. In the *IKR3_BSL_TT_HW* experiment, we employ the ln_expB2 model and the *I_comb* index to retrieve two thousand documents per query. Then, using the *I_in* and the *I_ex* indices and the aggregation model with $weight\_main = .1$, $weight\_in = .4$, and $weight\_in = .5$ we re-rank the top-1000 documents. These weights are selected manually based on the intuition that clinical trials with high similarity to the exclusion section should be heavily penalized. Following a similar intuition, the *IKR3_BSL_TT_MW* experiment also re-ranks the top-1000 documents of the *IKR3_BSL*. However, in this case, the selected weights are $weight\_main = .23$, $weight\_in = .33$, and $weight\_in = .44$. In this case, the system considers almost equally the importance of the similarity to a trial's main fields and inclusion criteria. However, again it penalizes those trials for which the patient has a high similarity to their exclusion criteria.

Lastly, in the *IKR3_TT_BW* and the *IKR3_TT_MW* experiments, we estimate the three topical relevance scores using the *I_main*, *I_in*, and *I_ex* for every document in the collection, and then we aggregate these scores. In the *IKR3_TT_BW* experiment, the weights were found via an exhaustive search conducted using last year's queries; as optimal weights were selected those that optimized P@10 measure. For the *IKR3_TT_MW*, the weights were equally allocated to the three similarity scores, i.e., equal to .33.

Table 1. Overall comparison with the TREC's median values.

	NDCG@10	PREC@10	Reciprocal Rank
TREC's Median	.392	.258	.411
IKR3_BSL	**.415**	.282	.529
IKR3_BSL_TT_HW	.352	.254	.506
IKR3_BSL_TT_MW	.405	**.288**	**.539**
IKR3_TT_BW	.395	.286	.513
IKR3_TT_MW	.408	.286	.535

4 Results

Table 1 presents the results obtained from the five submitted runs. A general observation is that none of the implemented experiments outperforms the baseline run (*IKR3_BSL*) in terms of NDCG@10. That is a reasonable outcome as the proposed methods aim at improving the task of finding eligible clinical trials for a patient; therefore, they penalize possible "excluded" trials. Specifically, the contribution of the similarity to a trial's exclusion criteria is always regarded negatively by the models. That is why one observes low NDCG scores but, at the same time, high precision-oriented scores. To improve the NDCG@10 measure, the model should consider the contribution of the similarity to a trial's exclusion criteria also positively, but with lower importance (i.e., $weight\_ex << weight\_in$ and $weight\_ex << weight\_main$). The highest PREC@10 and Reciprocal Rank values are obtained by the *IKR3_BSL_TT_MW* experiment in which the importance of the similarity obtained from the document parts that contribute positively to its overall relevance (i.e., similarity to main and inclusion parts) is almost equal to the importance of the one that contributes negatively ($weight\_main = .23$, $weight\_in = .33$, and $weight\_in = .44$). This finding suggests that one should consider the impact of the similarity to the main and inclusion parts. That is also supported in the *IKR3_BSL_TT_HW* experiments in which the positive contribution coming from the main document parts is almost neglected, and the relevance estimation relies on the similarity to the inclusion and the exclusion parts. Lastly, the *IKR3_TT_BW* and *IKR3_TT_MW* experiments suggest that the aggregation schema can be employed directly for document ranking, as the obtained results are close to the baseline.

5 Conclusions

This paper presents the results of our participation in the TREC 2022 Clinical Trials Track. Our objective was to investigate the effectiveness of our decision-theoretic model by heavily penalizing those clinical trials with which the patient has high topical similarity to their exclusion criteria. We experimented with weights that indicate how hard or soft the model should penalize those trials in two retrieval approaches, a full ranking and a re-ranking. The proposed methodology works similarly in both of them; however, its effectiveness is highly related to the selected weights. In conclusion, as the re-ranking approach yields greater performance, we aim to investigate the impact of the re-ranking depth, along with the selected weights.

6 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. Georgios Peikos, Oscar Espitia, and Gabriella Pasi. Unimib at trec 2021 clinical trials track, 2022.
2. Ching-Lai Hwang and Kwangsun Yoon. *Multiple Attribute Decision Making: Methods and Applications - A State-of-the-Art Survey*, volume 186 of *Lecture Notes in Economics and Mathematical Systems*. Springer, 1981.

3. Craig Macdonald and Nicola Tonellotto. Declarative experimentation ininformation retrieval using pyterrier. In *Proceedings of ICTIR 2020*, 2020.
4. Gianni Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.