DoSSIER at TREC 2023 Clinical Trials Track

Wojciech Kusa
wojciech.kusa@tuwien.ac.at
TU Wien
Vienna, Austria

Patrick Styll
patrick.styll@tuwien.ac.at
TU Wien
Vienna, Austria

Maximilian Seeliger
maximilian.seeliger@tuwien.ac.at
TU Wien
Vienna, Austria

Oscar E. Mendoza
o.espitiamendoza@campus.unimib.it
University of Milano-Bicocca
Milan, Italy

Allan Hanbury
allan.hanbury@tuwien.ac.at
TU Wien
Vienna, Austria

ABSTRACT

This paper describes the experimental setup and results of the DoSSIER team's participation in the Clinical Trials Track at TREC 2023. The primary objective of this track was to identify clinical trials for which patients meet the eligibility criteria. Our approach uses pipeline-based models, including large language models (LLMs) for query expansion and entity extraction techniques to augment both queries and documents. In our pipelines, we tested two different first-stage retrieval models, followed by a neural re-ranking framework that leverages topical relevance and eligibility criteria. We add to the pipeline a GPT-3.5-based question-answering post-processing step. Our findings demonstrate that the neural re-ranking and subsequent LLM post-processing notably enhanced performance. Future research will focus on a comprehensive assessment of the impact of query and document representation strategies on retrieval efficacy.

KEYWORDS

TREC Clinical Trials, clinical trials matching, neural re-ranking, TCRR, information extraction

1 INTRODUCTION

This paper presents an overview of the DoSSIER team's submissions to the TREC 2023 Clinical Trials (CT) Track[1]. The DoSSIER group consisted of members of the IR group of TU Wien and the University of Milano-Bicocca. We continued our focus from our previous TREC participation on efficient approaches for CT retrieval (CTR) and considered domain-specific characteristics of the retrieval task, neural re-ranking and information extraction methods.

Specifically, we follow a pipeline-based approach. We employed different first-stage and neural re-ranking models and compared how to enhance queries and documents with extracted information in order to increase the ranking performance. We also investigate the results of the final post-processing using large language models (LLMs). For the first stage retrieval, we enrich the textual representation of the clinical trial by extracting different elements such as keywords, entities, and sections, and compare BM25 retrieval based on different input text of the clinical trial [3]. We evaluate two different first-stage retrieval models, BM25 and DFR, implemented in two interfaces: ElasticSearch and PyTerrier. We fine-tune a cross-encoder on different parts of the clinical trials following the topical-criteria re-ranking (TCRR) model from Kusa et al. [3] and employ it for the re-ranking of two base runs. Finally, we test GPT-3.5-based post-processing using a question-answering approach in a zero-shot setting.

Within our experiments on the Clinical Trials track, we want to investigate the following research questions:

- **RQ1** How does the choice of the first-stage retrieval model influence the CTR precision?
- **RQ2** What is the impact of the TCRR neural re-ranking model on finding CT for which a patient is eligible?
- **RQ3** Can prompting an LLM (in a zero-shot scenario) help post-processing early-stage ranks and further improve CTR?

The evaluation results show that TCRR exhibits consistent performance improvements compared to the first stage retrieval models. We have also seen improvements in the first-stage using the improved entity extraction and normalisation pipeline. Furthermore, we reach the highest effectiveness in terms of reciprocal rank with the run post-processed by the GPT-3.5 model, suggesting that LLMs can contribute positively to CTR in a zero-shot setting.

2 METHODOLOGY

In this section, we give the background to our methodology. We describe the retrieval models used in the multiple stages of our pipelines, as well as the extraction methods used for query and document representations in CTR.

2.1 Query reformulation

Topics for the TREC CT 2023 edition are given in the format of disease "templates" with structured key information. This schema differs from the one followed in the 2021 and 2022 editions, which considered free text admission notes instead. We reformulate queries such that templates approximate free text admission notes in order to use our previous approaches [2, 3]. For that, we formulate a prompt to GPT-3.5 and use the output as topics (input prompt and example output are presented in Appendix A).

2.2 Query and document enhancement with entity extraction

We followed two different query and document enhancement approaches: (1) a two-stage lexical retrieval with a model-based feature extraction method, and (2) a query and document enrichment method as proposed by Kusa et al. [3]. In this section, we introduce the method (1). For further details on approach (2), refer to Kusa et al. [3].

[1] https://www.trec-cds.org/2023.html

For (1), we use the keywords provided for each topic to pre-select a total of 2000 trial documents in the domain of the topic keyword using the BM25 model.

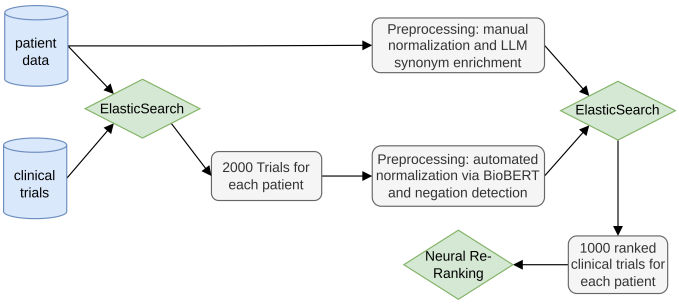

Figure 1: Overall Pipeline Architecture.

As a next step, we create a pipeline (see Figure 1) to extract a structured form of the inclusion and exclusion criteria. This is achieved by first splitting them into sentences, using the `spacy` English pretrained pipeline. Then, we use two different BioBERT fine-tunings for named entity recognition. One of them is fine-tuned for diseases, and the other one is specialized in chemicals and therefore also many terms relating to medication [1]. As a last step in the pipeline, we look for negations utilising yet another BERT fine-tuning optimised for negations in the clinical domain [5]. If a negation is recognized on a certain criterion, it is moved to the opposing category (i.e. a negated inclusion criterion is moved to the exclusion criteria in its non-negated form).

The small number of topics allows for a manual pre-processing of the patient data to match the described representation of exclusion and inclusion criteria. Furthermore, we used LLM-supported synonym enrichment to broaden the possible lexical matches with inclusion and exclusion criteria. In detail, we prompted the GPT-3.5 API to provide normalized synonyms for medical terms, which were manually selected from the topics given to us, in order to expand the semantic richness of our data. This process involved formulating a prompt that would efficiently leverage the deep contextual understanding of GPT-3.5, thereby surpassing traditional manual methods in both speed and scope (details on input prompt in Appendix B). Figure 2 presents an example of the normalization procedure.

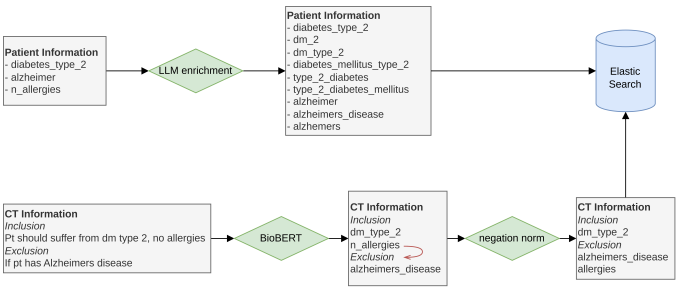

Figure 2: Example of normalization.

Table 1: Summary of our submitted TREC CT 2023 runs.

Run name	Description
DoSSIER_1	Transformer models fine-tuned on clinical data used to standardise Clinical Trials and topics. Topics were enhanced with understanding through Synonym Enrichment using Language Models (LLMs). BM25 was implemented in ElasticSearch to retrieve relevant results and devised a custom query, incorporating specialised analyzers within the ElasticSearch mappings, to match the normalised data effectively.
DoSSIER_2	Topics reformulated with GPT-3.5. DFR retrieval model from PyTerrier. Both documents and queries expanded using the approach from Kusa et al. [3] for drug and disease mentions in past, current and family medical conditions.
DoSSIER_3	Neural re-ranking with the TCRR approach using the BlueBERT model of the output from *DoSSIER_1* run. Re-ranking for the top 100 retrieved documents for each patient.
DoSSIER_4	Neural re-ranking with the TCRR approach using the BlueBERT model of the output from *DoSSIER_2* run. Re-ranking for the top 200 retrieved documents for each patient.
DoSSIER_5	Post-processesed *DoSSIER_4* run with the GPT-3.5 model using question-answering based on the eligibility criteria section. Filtering for each patient is done until the first ten of them are included or the number of excluded is equal to 50.

2.3 First-stage retrieval

We tested two different first-stage retrieval models implemented in two interfaces. On the one hand, we used BM25, which was implemented in ElasticSearch. On the other hand, we tested the DFR model from the PyTerrier Python library.

2.4 Neural re-ranking

For the neural re-ranking, we used the TCRR approach [3]. We chose the BlueBert [4] as the base model[2]. We fine-tuned the model on the 2021 Clinical Trials dataset as described in Kusa et al. [3].

2.5 Post-processing

Finally, we used the `gpt-3.5-turbo-0613` model using the OpenAI API[3] for post-processing of retrieval results. We used the question answering approach where the model is asked to make a binary *'yes'*/*'no'* decision based on the eligibility criteria section, concerning the diagnosis and current medical condition of the patient. We set our total experimental budget to 20.00 USD (0.50 USD per patient).

Table 2: Official TREC CT 2023 evaluation results.

Measure	nDCG@5	nDCG@10	Prec@10	Recip Rank
DoSSIER_1	0.6271	0.6238	0.4757	0.6268
DoSSIER_2	0.6347	0.6389	0.3622	0.5193
DoSSIER_3	0.6653	**0.6837**	**0.5838**	0.6421
DoSSIER_4	**0.6820**	0.6646	0.4108	0.6234
DoSSIER_5	0.6689	0.6718	0.4730	**0.6546**
TREC median	—	0.6484	0.3937	0.5381

3 EXPERIMENT AND RESULTS

3.1 Official submissions

The evaluation of the TREC Clinical Trials 2023 (TREC CT 2023) was aimed at assessing various information retrieval models, with a particular focus on their efficacy in clinical trial search tasks. A summary of our submitted runs is presented in Table 1. The official evaluation results, as presented in Table 2, delineate the performance of our submitted runs across nDCG at 5 and 10, Precision at 10, and Reciprocal Rank.

In addressing **RQ1**, which queries the influence of the first-stage retrieval model on the outcomes, we observe notable discrepancies when comparing *DoSSIER_1* and *DoSSIER_2*. *DoSSIER_1* demonstrated superior performance in both Prec@10 and RR relative to *DoSSIER_2*, suggesting that the retrieval model employed in *DoSSIER_1* is more adept at ranking relevant trials at the top of the results. On the other hand, the nDCG@5 and nDCG@10 of *DoSSIER_2* surpasses that of *DoSSIER_1*, indicating that *DoSSIER_2* was better overall in retrieving topically relevant trials.

RQ2 examines the impact of the TCRR neural re-ranking model on identifying eligible trials for patients. The comparative analysis between *DoSSIER_1* and *DoSSIER_3*, as well as *DoSSIER_2* and *DoSSIER_4*, reveals that *DoSSIER_3*, which integrates the TCRR model, outperforms other runs, especially in terms of nDCG@10 and Prec@10. This underscores the efficacy of the TCRR model in refining search results.

Lastly, **RQ3** evaluates whether post-processing with large language models in a zero-shot approach can further refine results. The juxtaposition of *DoSSIER_4* and *DoSSIER_5* indicates marginal enhancements, such as an improved reciprocal rank and P@10 for *DoSSIER_5*. However, these improvements are not as substantial as those witnessed with the implementation of the TCRR model. It is anticipated that with a more substantial experimental budget and the application of in-context learning, this strategy could yield more pronounced improvements.

In conclusion, the inclusion of the TCRR neural re-ranking model (*DoSSIER_3* and *DoSSIER_4*) has evidenced the strongest influence on the precision of clinical trial searches, markedly enhancing the relevancy of the results. Conversely, while the zero-shot approach using large language models did yield some improvements, it did not produce the considerable gains observed with the TCRR model, hinting at its potential limitations within this domain.

[2]https://huggingface.co/bionlp/bluebert_pubmed_uncased_L-12_H-768_A-12
[3]https://platform.openai.com

3.2 Ablation studies

We have experimented with various aspects of this pipeline after the TREC CT 2023 relevance judgments were published. The experiments primarily deal with the concept of LLM synonym enrichment, stemming and query boosting in ElasticSearch. The focus lies on improvements or behavioural aspects of both *DoSSIER_1* and *DoSSIER_3* (for neural reranking purposes).

3.2.1 Effect of LLM synonym enrichment. The comparative analysis of results in Table 3 underscores the important role of synonym enrichment in enhancing the performance of our approach, as detailed in section 2.2. The extreme contrast in performance metrics between the runs — with and without synonym enrichment — clearly illustrates the value added by including synonyms. The *DoSSIER_1* run, which includes synonym enrichment, shows significantly higher scores in all evaluated metrics, including nDCG and Precision. This enhancement can be attributed to the broader coverage and deeper understanding of the clinical domain that synonym enrichment offers. By extending patient data with synonyms, the approach becomes more robust in capturing the varied linguistic expressions of the same medical concepts, therefore increasing the likelihood of accurately matching patient profiles with relevant clinical trials. This enrichment is especially crucial in a field like medicine, where terminological variation is relatively high. In general, these results undermine synonym enrichment as a key component in the effectiveness of our approach.

Table 3: LLM synonym enrichment measures.

Measure	nDCG@5	nDCG@10	Prec@10	Recip Rank
no synonyms	0.2879	0.2666	0.1892	0.3684
DoSSIER_1	**0.6271**	**0.6238**	**0.4757**	**0.6268**

3.2.2 Effect of stemming. The results presented in Table 4 suggest that the application of stemming in the preprocessing of clinical trial data and patient information has a negligible impact on the performance metrics such as nDCG and Precision. This observation can be attributed to the already high level of standardization performed in clinical trial data and patient data, reducing the variability and language complexity stemming aims to address. It is also possible that the domain-specific nature of these texts makes common stemming techniques, which are generally designed for general language processing, less effective. Future work might explore more advanced natural language processing techniques, such as domain-specific lemmatization or semantic-based processing, to further enhance the matching process.

Table 4: Stemming measures.

Measure	nDCG@5	nDCG@10	Prec@10	Recip Rank
no stemming	0.6267	**0.6238**	**0.4784**	**0.6305**
DoSSIER_1	**0.6271**	**0.6238**	0.4757	0.6268

3.2.3 Effect of query boosting. The experimentation with different query boosting strategies in *DoSSIER_1*, as outlined in Table 5, provides some insightful results about the impact of varying the emphasis on different aspects of clinical trial data. Our initial assumptions for *DoSSIER_1* favored the inclusion criteria, boosting their significance by 300%, while halving the weight of other elements of the clinical trial data, such as the title and summary. However, as table 5 suggests, the empirical results challenge this hypothesis. The *no boosting* run surprisingly outperforms all our submissions in terms of nDCG and Precision, indicating that a more balanced approach proves to be beneficial. This finding suggests that the inclusion criteria, while important, should not overshadow other relevant aspects of the trials. The *inclusion boosting* run, although effective, was outperformed by the neutral approach, which poses strong evidence that we underestimated the impact of the other criteria. Conversely, the *others boosting* scenario, which enhanced the weight of non-inclusion criteria, presented an interesting outcome. Particularly noteworthy is the significant improvement in reciprocal rank, implying that these other aspects of the trials hold important relevance in aligning patient profiles with suitable clinical trials. This unexpected incline in performance highlights the complexity of information retrieval challenges in the clinical domain, where multiple factors simultaneously influence the relevance and suitability of trials for patients. These observations show the importance of refining query formulation, balancing various elements to optimize matching accuracy.

Table 5: Query Boosting measures.

Measure	nDCG@5	nDCG@10	Prec@10	Recip Rank
no boosting	**0.6895**	**0.6775**	**0.5351**	0.7063
inclusion boosting	0.6538	0.6534	0.5162	0.6427
others boosting	0.6584	0.6467	0.5270	**0.7167**
DoSSIER_1	0.6271	0.6238	0.4757	0.6268

4 CONCLUSION

Our participation in the Clinical Trials Track at TREC 2023 has yielded promising results, substantiating the efficacy of neural re-ranking and LLM-based post-processing in improving patient eligibility matching for clinical trials. Our approach integrated advanced techniques such as large language model-based query expansion and entity extraction to enrich the search process. Despite the advancements demonstrated, the impact of different query and document representations remains an open question, meriting further investigation. Our future work will concentrate on understanding the nuances of representation to further enhance retrieval performance.

ACKNOWLEDGMENTS

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] Álvaro Alonso Casero. 2021. *Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature.* Ph. D. Dissertation. ETSI_Informatica.

[2] Wojciech Kusa and Yasin Ghafourian. 2021. DOSSIER at TREC 2021 Clinical Trials Track. *TREC 2021* (2021).

[3] Wojciech Kusa, Oscar E. Mendoza, Petr Knoth, Gabriella Pasi, and Allan Hanbury. 2023. Effective Matching of Patients to Clinical Trials using Entity Extraction and Neural Re-ranking. *Journal of Biomedical Informatics* JBI (2023).

[4] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. arXiv:1906.05474 [cs.CL]

[5] Betty van Aken, Ivana Trajanovska, Amy Siu, Manuel Mayrdorfer, Klemens Budde, and Alexander Loeser. 2021. Assertion Detection in Clinical Notes: Medical Language Models to the Rescue?. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations.* Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2021.nlpmc-1.5

A CONVERSION TO FREE-TEXT FORMAT

Input prompt:

```
Consider this kind of format:

<topic number="40" template="type 2 diabetes">
<field name="definitive diagnosis">yes</field>
<field name="HbA1c">6.3</field>
<field name="glucose">115 fasting blood sugar</field>
<field name="BMI">40</field>
<field name="insulin">no</field>
<field name="metformin">8.5 mL</field>
<field name="other anti-diabetic drugs">no</field>
<field name="diet restrictions">low-calorie</field>
<field name="exercise">no</field>
<field name="ketoacidosis history"/>
<field name="comorbidities">chronic kidney disease</field
    >
<field name="hospitalization events">never</field>
</topic>

Convert this into free text. Take this style as an
    example:

Patient is a 45-year-old man with a history of anaplastic
     astrocytoma of the spine complicated by severe
    lower extremity weakness and urinary retention s/p
    Foley catheter, high-dose steroids, hypertension,
    and chronic pain. The tumor is located in the T-L
    spine, unresectable anaplastic astrocytoma s/p
    radiation. Complicated by progressive lower
    extremity weakness and urinary retention. Patient
    initially presented with RLE weakness where his
    right knee gave out with difficulty walking and
    right anterior thigh numbness. MRI showed a spinal
    cord conus mass which was biopsied and found to be
    anaplastic astrocytoma. Therapy included field
    radiation t10-l1 followed by 11 cycles of
    temozolomide 7 days on and 7 days off. This was
    followed by CPT-11 Weekly x4 with Avastin Q2 weeks/
    2 weeks rest and repeat cycle.

Do NOT by any terms make up information and stay short.
    If age and gender are not given in the xml format,
    stay neutral and do not make up age or gender.
```

Example Output:

```
The patient, diagnosed with type 2 diabetes (HbA1c: 6.3%,
     fasting blood sugar: 115), has a BMI of 40. Insulin
     is not currently prescribed, but they are taking
    8.5 mL of metformin. No other anti-diabetic drugs
    are being used. The patient follows a low-calorie
    diet and does not engage in regular exercise. There'
    s no history of ketoacidosis, but they do have
    chronic kidney disease as a comorbidity.
    Hospitalization due to diabetes has not occurred.
```

B LLM-SUPPORTED SYNONYM ENRICHMENT

We use the following prompt for the LLM-supported synonym enrichment:

```
Provide as many synonyms as possible (at least 5) for
    each word. Use medical synonyms only. Stay in the
    same format (lower case, use _ instead of
    whitespaces). Provide the answer as a String array.

"type_2_diabetes","6.3%_HbA1c", "fasting_glucose_of_115",
     "40_BMI", "8.5mL_metformin", "diet_restrictions", "
    low-calorie_diet", "comorbidities", "
    chronic_kidney_disease"
```