WisPerMed Text at TREC Clinical Trials Track 2021

Henning Schäfer[1,2,*], Ahmad Idrissi-Yaghir[2,*], Wolfgang Galetzka[3,*], Marie Bexte[4,*], and Christoph M. Friedrich[2,3]

[1] Institute for Transfusion Medicine, University Hospital Essen, Hufelandstraße 55, 45147 Essen, Germany
[2] Department of Computer Science, University of Applied Sciences and Arts Dortmund (FHDO), Emil-Figge Str. 42, 44227 Dortmund, Germany
[3] Institute of Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Hufelandstraße 55, 45147 Essen, Germany
[4] Language Technology Lab, University of Duisburg-Essen
Lotharstraße 65, 47057 Duisburg, Germany
henning.schaefer@uk-essen.de, ahmad.idrissi-yaghir@fh-dortmund.de, wolfgang.galetzka@uk-essen.de, marie.bexte@uni-due.de, christoph.friedrich@fh-dortmund.de

Abstract. This paper describes the submissions of the WisPerMed Text group to the TREC Clinical Trials Track 2021. It aims to overcome the problems in patient recruitment that often lead to delays or even discontinuation of clinical trials. The focus here is finding methods to improve the process of matching patient case descriptions to eligible clinical trials. For this purpose, different systems were proposed and tested to rank the trials for each patient topic. These systems utilize methods such as transformer-based models, BM25 and keyword extraction. Additionally, Unified Medical Language System (UMLS) was used in an attempt to find relevancy between patient topics and clinical trials based on biomedical concepts. The results obtained showed that the BM25 model based on keyword extraction performed the best out of all our submissions.

1 Introduction

Evidence-based medicine relies on clinical trials to translate research findings into practice at the point of care. New treatments, whether for cancer care or the introduction of new drugs, all rely on running through the various phases of clinical trials. The big challenge here is to recruit enough participants for these trials [1]. A major problem in doing so is to match patients to the inclusion and exclusion criteria of the trial. Until now, clinicians have to be constantly aware of clinical trials relevant to their field and assess the patient eligibility based on the semi-structured electronic health records (EHRs) [2]. Automated systems can help reduce the time spent searching for suitable trials by using existing data from the EHR [3], ideally bringing up options that would have otherwise remained under the radar. This would be beneficial both for patients and clinicians, as it will improve the recruitment process and prevent trials from being discontinued due to the lack of patients. The TREC Clinical Trials Track 2021[5] challenge is an important step in this direction, as it aims to identify relevant trials based on patient records. Here, 75 synthetic patient descriptions are provided in the form of topics, for which eligible clinical trials are to be ranked.

In this work, some classical information retrieval options are compared with current transfer learning and transformer-based methods. Since the research area here is just being explored and there is little annotated data on clinical trials, the investigation of transfer learning approaches is the subject of study here. The task is complicated by the fact that both the criteria and the background of the patient given by the topics must match the trials. In practice, when using an automated search for eligible trials, there is no point in matching subjects to a trial where all criteria are objectively fulfilled, but the topic refers to a different pathology.

Therefore, the approaches are twofold, on the one hand a preprocessing takes place for each topic, which should filter the selection of trials on the basic criteria of age and gender. On the other hand, from this sub-selection, trials will be selected and sorted, e.g. via classical approaches such as BM25 or via semantic similarities determined from pre-trained biomedical models. The focus of this approach is

[*] The authors contributed equally to this work
[5] https://www.trec-cds.org/2021.html, last accessed: 02.11.2021

also to include the background of the topics for matching trials, at the expense of precision of meeting all inclusion and exclusion criteria.

The remainder of this paper is structured as follows. First, the available data set for the challenge and the preprocessing steps are described. The methodology section is then divided into the five individual submissions. For each submission, details on the implementation are given. The results of the submissions are evaluated and compared with all TREC submissions (median and best). Finally, discussion, conclusion and future works classify the results and main findings and show potential weaknesses of the approaches, as well as ways to improve them in the future.

2 Data

For this task, a collection of 375,580 publicly accessible clinical trials descriptions was used, which was obtained from `https://clinicaltrials.gov` on April 27, 2021. In addition, 75 topics were made available, which consist of synthetic patient cases created by individuals with medical training. These topics simulate an admission statement in an EHR of a patient.

The clinical trials corpus consists of different fields that describe the trials and can be used for patient recruitment. These include a brief summary of the trial, eligibility criteria, gender, minimum and maximum age, etc. Out of all trials, 841 were labeled as "trials of device that is not approved or cleared by the U.S. FDA" and all their fields did not contain any information. Therefore, these trials were removed from the corpus. The remaining trials also had empty fields in some cases. For the selected fields in work, the number of null values is shown in Table 1.

Table 1. Number of empty fields in trials corpus

Field	Number of null values
url	0
#nct_id	0
title	0
official_title	10085
brief_summary	1
detailed_description	125187
conditions	19
eligibility_criteria	91
gender	21
min_age	30110
max_age	179020

Since many trials aren't provided with an official title and a detailed description, these fields were not considered for this work.

The provided topics are unstructured text documents, which have a length that varies between 5 and 10 sentences. These topics can be considered as the query to search for suitable clinical trials for each patient case.

3 Preprocessing

As a preprocessing step, age and gender information were extracted from the topics using regular expressions. For this purpose, the standard *re*[6] module in python was used. To handle the different units, i.e. days, weeks, months and days, in which the age was specified, were converted into days. The extracted age and gender information was then used to restrict the trials to a set of trials, for which a patient is eligible. To do this, we used the fields `min_age`, `max_age` and `gender` of a trial. Any trial that did not contain sufficient information to exclude a patient from participating were kept as candidate trials for the respective patient.

[6] `https://docs.python.org/3/library/re.html`, last accessed: 10.02.22

For most of the used approaches, no text preprocessing was utilized except for the third submission, where punctuation was removed, and the documents were tokenized.

4 Approaches

This section describes the approaches used for each submission. First, a ranking based on transformer-based document embeddings and cosine similarity was used. The second submission consists of keyword extraction and BM25 document ranking. For the third submission, the ranking is based on a combination of transformer-based embeddings and TF-IDF features. In the fourth submission, a new ranking was performed using transformer-based embeddings for a ranking retrieved using Elasticsearch. Whereas, the final ranking was based on a score considering the number of normalized medical terms appearing in the title of the studies and in the description of the patients, as well as the prioritization of rare terms.

4.1 Submission 1

For this submission, a semantic search was preformed. First, the brief summaries of each available clinical trial and the different topics were embedded using Sentence-Transformers[7] [4]. Here, a clinical-oriented pre-trained BERT [5] model was utilized. This model Clinical BioBERT [6] was initialized from BioBERT [7] and trained on additional clinical text from approximately 2 million notes in the MIMIC-III v1.4 database [8]. To encode the text to embeddings, the default parameters were used except for the *max sequence length*, which was increased to 500 tokens. Afterwards, the cosine similarity between the embeddings of the topics and those of the brief summaries was computed. Based on the obtained similarity scores, the clinical trials for each topic were ranked.

4.2 Submission 2

In this approach, the Clinical BioBERT model was this time utilized with KeyBERT[8] [9], which uses the document embedding and word embeddings of a specific document to find the words that are the most similar to the document based on cosine similarity. These words could then be considered as the keywords that best describe the entire document. Using KeyBERT the top 10 keywords were extracted from each Topic. Here, the parameter *ngram range* was set to (1,1) so that only unigrams are selected and the *stop words* parameter was set to English, which prompts the extraction method to remove English stop words from the document.

To rank the clinical trials, the Okapi BM25 [10] was used, which is a retrieval function to estimate the relevance of documents to a given query based on the query terms appearing in each document [11]. Here, the implementation of BM25 in Rank-BM25[9] was used with the parameter k_1 set to 1.5 and b set to 0.75. The variable k_1 is a tuning parameter, which is used to limit the influence of a single search term on the score of a document. Whereas, b is used to control the effect of the document length on the score. The selected keywords were used as the search query and the brief summaries of the clinical trials as the documents to search from. The scores for each query were then calculated and based on them, a ranking of the clinical trials was identified for each individual topic.

4.3 Submission 3

For this submission, BioBERT was used to encode the topics and the combination of the title and the brief summary of the clinical trials. In addition to the dense BioBERT embeddings, a weighted similarity based on sparse representation was included in the final similarity calculation [12]. The semantic match between mention domain (topics) and candidates domain (clinical trials), as well as the similarity based on character level representations, were incorporated into this ranking.

The dense representation is described in Equation 1.

[7] `https://github.com/UKPLab/sentence-transformers` (Version 1.2.0), last accessed: 02.11.2021
[8] `https://github.com/MaartenGr/KeyBERT` (Version 0.4.0), last accessed: 02.11.2021
[9] `https://github.com/dorianbrown/rank_bm25` (Version 0.2.1), last accessed: 02.11.2021

$$e_d^m = BioBERT(\overline{m})[CLS] \in \mathbb{R}^h \tag{1}$$

Here, $\overline{m} = (\overline{m}_1, \ldots, \overline{m}_l)$ is a sequence of subtokens of the topic m by the WordPiece tokenizer [13]. Parameter h denotes the hidden dimension ($h = 768$). Special token [CLS] denotes the token that BERT-style models use to compute a single representative vector of an input.

Similarly, the title and brief summary dense representation of all candidate trials is calculated (Equation 2).

$$e_d^n = BioBERT(\overline{n})[CLS] \in \mathbb{R}^h \tag{2}$$

With $\overline{n} = (\overline{n}_1, \ldots, \overline{n}_l)$ being the sequence of a candidate trial.

$$S_{dense} = f(e_m^d, e_n^d) \tag{3}$$

Here, e_m^d and e_n^d denote the BERT embedded representation of a topic and a candidate trial, where f can be any similarity function.

The sparse representation is obtained through TF-IDF [14] and is calculated based on the character-level ngrams statistics computed over all candidates $n \in N$. Equation 4 shows the resulting sparse representation.

$$S_{sparse} = f(e_m^t, e_n^t) \tag{4}$$

Here, e_m^t and e_n^t denote the TF-IDF representation of a topic and a candidate trial, where f can likewise be any similarity function. For both dense and sparse similarity, experiments were conducted using Maximum Inner Product Search (MIPS) as proposed by [12]. Experiments with euclidean distance and cosine similarity could not be conducted due to time constraints and higher computational cost.

The final similarity function to rank trials is shown in Equation 5.

$$S(m, n) = S_{dense}(m, n) + \lambda S_{sparse}(m, n) \in \mathbb{R} \tag{5}$$

Here, function $S(m, n)$ indicates both similarities between an input topic m and a candidate trial n. Parameter λ is a trainable scalar weight for the sparse score used for balancing both similarities. The λ parameter was optimized by using the iterative candidate retrieval proposed by [12]. They used the National Center for Biotechnology Information (NCBI) disease corpus to train the parameter and weight between dense and sparse representations. During training of λ, whenever dense and sparse candidates for a disease mention overlap, more dense candidates were added to match up to the number of sparse candidates by changing λ. This resulted in $\lambda = 1.4$.

According to [12] this helps for two reasons: First, it challenges the top candidates with more difficult negative candidates, which helps to get a more accurate dense representation. Second, it increases the chances of finding previously unseen positive patterns in the top candidates.

Originally intended for normalizing concepts such as diseases, the parameter was adopted for this work because it was trained for the biomedical domain and there is not enough annotated data to train the parameter for the task of ranking trials to topics.

4.4 Submission 4

In this approach, an initial subset of trials was obtained by using Elasticsearch[10] Lucene [15] BM25 [10] for a first-stage retrieval. The results of this ranking were then re-ranked based on the transformers approach previously described in Section 4.3. For the initial retrieval, the parameter k_1 was set to 1.2, and b to 0.75. The re-ranking was performed by sorting the top 1000 trials from the initial retrieval using the scores of similarity between dense embeddings and sparse TF-IDF representation.

[10] `https://hub.docker.com/_/elasticsearch` (Version 7.14.1), last accessed: 02.11.2021

4.5 Submission 5

In the approach used for this ranking, medical conditions for which only few trials were available were prioritized. This was done in following steps.

First, medical terms appearing in the patients' descriptions and in the titles of the trials were extracted and normalized based on Unified Medical Language System (UMLS) using ScispaCy [16]. Here ScispaCy[11] was used with the model *en_core_sci_lg* and the UMLS *entitylinker*, which links terms to concepts from the UMLS 2020 AB release.

In the second step, a weight was defined for each medical term c as:

$$w_c := \frac{1}{\text{Number of trial titles containing } c}. \tag{6}$$

Afterwards, a first score of a trial with the set of medical terms C_t for a patient with a set of medical terms C_p was calculated as:

$$\text{Score of patient} := \sum_{c \in C_p \cap C_t} w_c. \tag{7}$$

This score was identical for many trials, as trial title and patient description often had at most one medical term in common. Thus, a second score was built and applied for trials for which the first score (7) was identical.

To this end, inclusion and exclusion criteria of the trials were extracted using a Clinical Trial Parser[12] [17] and normalized based on UMLS concepts. For a clinical trial, sets C_i and C_e of inclusion terms and exclusion terms were obtained this way. The share of fulfilled inclusion criteria,

$$\frac{|C_p \cap C_i|}{|C_i|}, \tag{8}$$

was used to rank trials with identical (7). Finally, trials with identical (7) and (8) were ranked according to the negative share of violated exclusion criteria, i.e. $-\frac{|C_p \cap C_e|}{|C_e|}$.

5 Results

For the evaluation, the trials were graded accordingly: an *eligible* trial was given a score of 2, an *excluded* trial a score of 1 and 0 to trials described as *not relevant*. A trial is eligible, when the trial is relevant and the patient is eligible for this trial. Trials are denoted as excluded, when the trial is relevant to the described condition, but the patient is not eligible for the trial due to the exclusion criteria. The mentioned grades were then used to compute the Normalized Discounted Cumulative Gain (NDCG). For metrics based on binary judgements such Precision and Reciprocal Rank, excluded trials were treated as not relevant.

The results of this work's submissions can be seen in Table 2, which shows the NDCG@5. NDCG@10, P@10 and RR scores of the different topics.

Based on obtained results, out of the five submitted runs, only the BM25 ranking based on keyword extraction achieved scores above the TREC median for automatic runs.

6 Discussion

Somewhat surprisingly, transformer-based models performed poorly compared to simple methods such as BM25. Here, we suspect that using off-the-shelf pre-trained language models might not achieve good results without any additional fine-tuning, especially in the context of semantic similarity.

Additionally, our approach using UMLS concepts had the drawback of ScispaCy returning a rather large set of concepts, while only a smaller subset of these concepts refers to the main concern regarding

[11] `https://github.com/allenai/scispacy` (Version 0.4.0), last accessed: 02.11.2021

[12] `https://github.com/facebookresearch/Clinical-Trial-Parser` (Commit: 424a952bf3927413db76995d07e0d69529e88337), last accessed: 02.11.2021

Table 2. Results of the presented runs in comparison to best runs. Bold numbers indicate the highest scores.

Submission	**NDCG**		**P@10**	**RR**
	@5	@10		
1 Clinical BioBERT + Cosine	0.0760	0.0763	0.0493	0.1000
2 KeyBERT + BM25	**0.3432**	**0.3336**	**0.2133**	**0.4069**
3 BioBERT + TF-IDF	0.2047	0.1948	0.1427	0.2863
4 Elasticsearch + BioBERT	0.1852	0.1583	0.0973	0.2630
5 UMLS	0.1449	0.1372	0.0840	0.2122
All TREC Submissions: Median	-	0.3040	0.1613	0.2942
All TREC Submissions: Best	-	0.8491	0.7480	1.0000
Best Team	-	**0.7118**	**0.5933**	**0.8162**

a patient or trial. To handle this, a weighting strategy assigning higher weights to these core concepts might be a viable adjustment when comparing trial and patient concepts. Another way of addressing this would be to introduce more preliminary filtering. While some general filtering of trials regarding age and gender of a patient was included, this could be widened to include preliminary filtering regarding the main concern of a patient. Ideally, this would lead to a much smaller amount of trials to rank and prohibit irrelevant trials that nonetheless share many of the patient properties from being ranked high.

Similar to the extraction of eligibility criteria using the Clinical Trial Parser, a promising approach would be to also extract a structured representation of criteria-relevant information mentioned in a topic description. This could facilitate an easier comparison between topic and trial and would lead to more explainable results.

7 Conclusion and future works

This work presented methods and results from the participation of the WisPerMed Text group in the TREC Clinical Trials Track 2021. The objective was to try different approaches, from traditional information retrieval methods to approaches that incorporate transformer-based models. Though most of the approaches did not perform well, they leave room for improvement.

One difficulty of this challenge was that the information is only available in unstructured topics. The matching of clinical trials with patients could be improved by taking structured and semi-structured patient data directly from the electronic health record. In addition, systems for formally reading in inclusion and exclusion criteria need to get better at detecting possible negations, as criteria end up in both sections depending on how they are worded.

8 Acknowledgement

This work was funded by PhD grants from the DFG Research Training Group 2535 Knowledge- and data-based personalization of medicine at the point of care (WisPerMed), University of Duisburg-Essen, Germany.

References

1. R. B. Gul and P. A. Ali, "Clinical trials: the challenge of recruitment and retention of participants," *Journal of Clinical Nursing*, vol. 19, no. 1-2, pp. 227–233, 2010.
2. B. G. Sully, S. A. Julious, and J. Nicholl, "A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies," *Trials*, vol. 14, no. 1, pp. 1–9, 2013.
3. P. M. Rothwell, "Commentary: External validity of results of randomized trials: disentangling a complex concept," *International journal of epidemiology*, vol. 39, no. 1, pp. 94–96, 2010.
4. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 11 2019.

5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.
6. E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly Available Clinical BERT Embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, (Minneapolis, Minnesota, USA), pp. 72–78, Association for Computational Linguistics, June 2019.
7. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
8. A. E. W. Johnson, T. J. Pollard, L. Shen, L. wei H. Lehman, M. Feng, M. M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, 2016.
9. M. Grootendorst, "Keybert: Minimal Keyword Extraction with BERT.," 2020.
10. G. Amati and C. J. Van Rijsbergen, "Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness," *ACM Transactions on Information Systems*, vol. 20, pp. 357–389, Oct. 2002. http://doi.acm.org/10.1145/582415.582416.
11. S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, pp. 333–389, 01 2009.
12. M. Sung, H. Jeon, J. Lee, and J. Kang, "Biomedical Entity Representations with Synonym Marginalization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 3641–3650, Association for Computational Linguistics, July 2020.
13. Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *CoRR*, vol. abs/1609.08144, 2016.
14. H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting TF-IDF Term Weights as Making Relevance Decisions," *ACM Transactions on Information Systems*, vol. 26, pp. 13:1–13:37, June 2008. http://doi.acm.org/10.1145/1361684.1361686.
15. B. Milosavljević, D. Boberić, and D. Surla, "Retrieval of bibliographic records using Apache Lucene," *The Electronic Library*, 2010.
16. M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, (Florence, Italy), pp. 319–327, Association for Computational Linguistics, Aug. 2019.
17. Y. Tseo, M. I. Salkola, A. Mohamed, A. Kumar, and F. Abnousi, "Information Extraction of Clinical Trial Eligibility Criteria," *CoRR*, vol. abs/2006.07296, 2020.