CogStack Cohort at TREC 2022 Clinical Trials Track

Jack Wu, Zeljko Kraljevic, Thomas Searle, Richard Dobson, Daniel Bean

Biostatistics & Health Informatic, King's College London, United Kingdom
{ho_chung.wu, zeljko.kraljevic, thomas.searle, richard.j.dobson, daniel.bean}@kcl.ac.uk

Abstract. This notebook paper describes the methodology used to produce the retrieval results we submitted for TREC 2022 Clinical Trials Track. The method is based on named entity recognition and linking (NER+L). Medical Concept Annotation Tool (MedCAT) is used to perform NER+L on the topics and documents to produce entities in the SNOMED Clinical Terms ontology. The clinical terms extracted by the annotation process are used for indexing and retrieval purposes. The retrieval model used is a passage-based retrieval model that gives different weights to different document portions.

Keywords: Clinical trials retrieval, Named Entity Recognition and Linking.

1 Introduction

Clinical trials are important research studies to evaluate and develop medical treatments and procedures. Patient recruitment is a main challenge in performing clinical trials [1]. Being able to recruit a minimum number of patients satisfying a set of criteria given by a clinical trial within a specific timeframe is crucial to the success of the trial. However, manual inspection of patient records to identify potential patients to be included in a trail is time consuming and is associated with high costs. It is estimated that more than 80% of clinical trials fail to enroll enough patients on time [2].

In recent years, natural language processing (NLP) and machine learning (ML) technologies have been applied to electronic health records for cohort identification (e.g., [3, 4, 5]). These methods examine the unstructured data (i.e., free text) from patients' records and match the patients' medical conditions with the criteria set by clinical trials. An advantage of using NLP and ML technologies is that a large number of records can be processed automatically, within a relatively short period of time. This increases the chance of identifying suitable patients for clinical trials.

The TREC 2022 Clinical Trials Track focuses on the task of matching patients' electronic health records with a collection of clinical trials from the ClinicalTrials.gov[1] dataset, both are in text format. There are 50 topics where each topic is a synthetic patient description simulating an admission statement. Each topic is 5-10 sentences long. The topics are created by individuals with medical training. For the document collection

[1] https://clinicaltrials.gov/

(clinical trials), there are over 375k documents in XML format. Each document contains the title, brief summary, detailed description, eligibility criteria, etc. of a clinical trial. The objective of the task is to develop a retrieval system which outputs a ranked list of documents (clinical trials) for each of the topics (patient descriptions). For a particular topic, each of the returned documents in the ranked list is classified as either *eligible* (meeting inclusion criteria and exclusion criteria do not apply), *excluded* (meeting inclusion criteria but excluded due to exclusion criteria) and *not relevant* (not meeting inclusion criteria).

Our submitted runs focus on using Medical Concept Annotation Tool (MedCAT) [6] to annotate the topics and documents for retrieval while a basic passage-based vector space retrieval model is used. MedCAT is an open-sourced toolkit[2] to extract medical concepts from texts in electronic health records and link them to ontologies such as SNOMED-CT [7]. MedCAT uses a novel concept disambiguation algorithm that learns clinical concept similarity via Word vector contexts and employs a Bidirectional-Long-Short-Term-Memory (Bi-LSTM) model for contextualisation. Correction of spelling mistakes is performed during the annotation process. A cohorting tool based on MedCAT is also available for download[3]. In our runs, the extracted concepts are used to match a topic to the trial descriptions with Term Frequency (TF) x Inverse Document Frequency (IDF) scores in a passage-based retrieval model.

2 Methodology

In the pre-processing step, five XML tags are located in each of the documents (clinical trial descriptions):

(1) <brief_title>;
(2) <brief_summary>;
(3) <detailed_description>;
(4) <eligibility>; and
(5) <mesh_term>.

The content (i.e., text) of a tag forms one passage except for the <eligibility> tag. Inclusion criteria and exclusion criteria are extracted from the <eligibility> tag using Regular Expression to form two passages (one for inclusion criteria and another one for exclusion criteria, if the exclusion criteria exist). The tag <mesh_term> can have multiple occurrences in a document, they are concatenated to form a single passage. As a result, there are a total of six passages in a document and they are annotated using MedCAT individually, all other sections in the document are discarded. Each topic is also annotated using MedCAT with the same model.

Both the topics and documents are treated as bag-of-concepts and retrieval is done using a vector space model with uniform weights for query concepts and TFxIDF weights for document concepts. Different weights are also assigned to the six passages in a document with the weight of the exclusion criteria passage being negative. The

[2] https://github.com/CogStack/MedCAT
[3] https://github.com/CogStack/CogStackCohort

weights are tuned using the topics and relevance judgements from the TREC 2021 Clinical Trials Track by optimizing the average precision at 10. The tuned weights are listed below (Table 1). The exclusion criteria passage receives a negative weight to decrease the document's score when a concept is matched in it.

Table 1. Weights of different passages in a document.

Passage in a document	Weight
Brief title	0.4
Brief summary	0.2
Detailed description	0.6
Inclusion criteria	0.8
Exclusion criteria	-0.2
Mesh terms	0.6

Five runs are submitted, all of them are automatic runs with the following descriptions (Table 2):

Table 2. Runs submitted.

Run Name	Description
Run1	Vector space model for retrieval. MedCAT model 1.
Run2	Vector space model for retrieval. MedCAT model 2.
Run3	Query expansion (top 20). MedCAT model 1.
Run4	Query expansion (top 10). MedCAT model 1.
Run5	Query expansion (top 10). MedCAT model 2.

3 Concepts Extracted from Topics and Documents

This section describes concepts obtained from some example topics in Table 3. The most common concepts extracted from the 375k documents are also shown in Table 4.

Table 3. Example topics and SNOMED concepts extracted.

Example Topic	SNOMED concepts extracted
A 67-year-old woman comes to the clinic due to recent episode of choking[1], dysphagia[2], and cough[3]. Her other medical problems include hypertension[4], dyslipidemia[5], and osteoarthritis[6]. She does not smoke[7] or use alcohol. She lives with her husband and she is able to do her own daily activities. She used to teach elementary school. Blood pressure is 135/80 mm Hg. The patient's breath smells bad. Other physical examinations are normal. A barium swallow study reveals an abnormality in the upper esophagus with an outpouching at the junction of the lower part of the throat and the upper portion of the esophagus.	1. Choking (finding) 2. Dysphagia (disorder) 3. Cough (finding) 4. Hypertensive disorder, systemic arterial (disorder) 5. Dyslipidemia (disorder) 6. Osteoarthritis (disorder) 7. Non-smoker (finding)
A 30-year-old man who is a computer scientist came to the clinic with the lab result stating azoospermia[1]. The patient is sexually active with his wife and does not use any contraception methods. They have been trying to conceive[2] for the past year with no success. The patient has a past medical history of recurrent pneumonia[3], shortness of breath[4], and persistent cough[5] that produces large amounts of thick sputum[6]. The patient had multiple lung infections[7] during childhood. He does not smoke[8], use illicit drugs or alcohol. The patient has no history of other medical conditions including allergies or any kind of surgery. On physical examination, the digits show clubbing. An ultrasound shows bilateral absence of the vas deferens, and FEV1 was 75% on the respiratory function test.	1. Azoospermia (finding) 2. Trying to conceive (finding) 3. Recurrent pneumonia (disorder) 4. Dyspnea (finding) 5. Persistent cough (finding) 6. Thick sputum (finding) 7. Infectious disease of lung (disorder) 8. Non-smoker (finding)
A 47-year-old woman comes to the clinic complaining of dizziness[1]. She also has occasional nausea[2] and ringing in her right ear. The patient also has difficulty hearing while holding her phone to the left ear, although hearing in her right ear is normal. The dizziness[3] improves spontaneously, and she feels fine between episodes. Past medical history is notable for hypothyroidism[4] and low vit D level, for which she is using Levothyroxine[5] and Vit D pearl. She does not use tobacco or drink alcohol. Physical examination shows sensorineural hearing loss[6] in the left ear. She has only one-man sexual partner and menopaused 2 years ago.	1. Dizziness (finding) 2. Nausea (finding) 3. Dizziness (finding) 4. Hypothyroidism (disorder) 5. Levothyroxine (product) 6. Sensorineural hearing loss (disorder)

Table 4. Most extracted SNOMED concepts from the collection.

SNOMED Concept	Document Frequency
Malignant neoplastic disease (disorder)	102,703
General treatment (procedure)	95,937
Neoplasm (morphologic abnormality)	49,457
Diabetes mellitus (disorder)	47,850
Hypertensive disorder, systemic arterial (disorder)	37,885
Pain (finding)	34,998
Myocardial infarction (disorder)	32,746
Cerebrovascular accident (disorder)	32,496
Viral hepatitis type B (disorder)	32,034
Biopsy (procedure)	29,379
Antibiotic (product)	24,645
Viral hepatitis type C (disorder)	23,582
Medicinal product acting as contraceptive (product)	23,419
Disorder of cardiovascular system (disorder)	23,378
Blood coagulation disorder (disorder)	22,340
Heart disease (disorder)	20,823
Depressive disorder (disorder)	20,115
Finding of tobacco smoking behavior (finding)	20,067
Substance abuse (disorder)	19,834
Disorder of liver (disorder)	19,793

4 Experiments and Results

Table 5 shows the results of the five submitted runs which reports NDCG@10, Precision@10 and Reciprocal Rank (RR). The results from the MedCAT model 1 are better than model 2. Model 1 has a larger training set than model 2.While both models are proprietary, they can be can be permissioned upon request. There are also public models available to download[4]. Surprisingly, the three runs with query expansion perform significantly worse than expected. Further investigations such as parameter turning need to be conducted. Overall, the result from Run1 is comparable with the TREC's median results.

[4] https://github.com/CogStack/MedCAT

Table 5. Results of the submitted runs compared with the overall median.

	NDCG@10	P@10	RR
Run1	0.3725	0.2480	0.5516
Run2	0.3386	0.2400	0.4611
Run3	0.0131	0.0060	0.0110
Run4	0.0160	0.0080	0.0097
Run5	0.0128	0.0080	0.0128
TREC Median	0.3922	0.2580	0.4114

5 Conclusions

The results of our submitted runs to TREC 2022 Clinical Trials Track are presented. While using a simple retrieval model, the results from Run1 suggest that MedCAT is effective in identifying medical concepts in patient summaries and clinical trial descriptions. Further investigation is needed for the decreased performance in query expansion. Future work includes experimenting with other retrieval methods such as learning-to-rank and developing of a web application for users to search for clinical trials using SNOMED concepts.

6 Acknowledgements

This work is supported by The King's British Heart Foundation (BHF) Centre of Research Excellence.

References

[1] Fogel, D. B. (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemporary clinical trials communications*, 11, 156-164.

[2] Desai, M. (2020). Recruitment and retention of participants in clinical studies: Critical issues and challenges. *Perspectives in Clinical Research*, 11(2), 51.

[3] Yuan, C., Ryan, P. B., Ta, C., Guo, Y., Li, Z., Hardin, J., ... & Weng, C. (2019). Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26(4), 294-305.

[4] Ling, A. Y., Kurian, A. W., Caswell-Jin, J. L., Sledge Jr, G. W., Shah, N. H., & Tamang, S. R. (2019). Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. *JAMIA open*, 2(4), 528-537.

[5] Tissot, H. C., Shah, A. D., Brealey, D., Harris, S., Agbakoba, R., Folarin, A., ... & Asselbergs, F. W. (2020). Natural language processing for mimicking clinical trial recruitment

in critical care: a semi-automated simulation based on the LeoPARDS trial. *IEEE Journal of Biomedical and Health Informatics*, 24(10), 2950-2959.

[6] Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., ... & Dobson, R. J. (2021). Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117, 102083.

[7] Stearns, M. Q., Price, C., Spackman, K. A., & Wang, A. Y. (2001). SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium* (p. 662). American Medical Informatics Association.