An approach to relevant clinical trials retrieving (clinical_trials team)

Mariia Fedorova

ElmDoc B.V.
maria.fjodorowa@gmail.com

Abstract. In this notebook paper an approach to retrieving relevant clinical trials for patients' unstructured descriptions (electronic health records, EHTs) is described.

1 Introduction

In this work, an approach to the TREC Clinical Trials track is presented. The methods used are medical named entities recognition, described in the section 2; a preliminary selection of clinical trials from a database, containing more than 300 000 examples, described in the section 3; three ranking algorithms, described in the section 4 and evaluation results for each run with error analysis in the section 5.

2 Named entities recognition

Three classes of named entities were parsed from both topics (patients' descriptions) and clinical trials criteria: Conditions (pathologies, syndromes, symptoms and special non-pathological conditions, e.g. pregnancy), Procedures and Drugs. Two methods for named entities recognition were used: a string matching to the UMLS [1] concepts, implemented in the QuickUMLS Python library[1] [5] and BioBERT[2] [3], fine-tuned on the CHIA corpus [2].

The following UMLS semantic types were used for different named entities classes:

- Conditions
  - T047 (Disease or Syndrome, ex.: Diabetes Mellitus; Drug Allergy; Malabsorption Syndrome)
  - T048 (Mental or Behavioral Dysfunction, ex.: Agoraphobia; Cyclothymic Disorder; Frigidity)
  - T020 (Acquired Abnormality, ex.: Hemorrhoids; Hernia, Femoral; Cauliflower ear)
- Procedures

[1] https://github.com/Georgetown-IR-Lab/QuickUMLS

[2] https://huggingface.co/dmis-lab/biobert-base-cased-v1.1

    - T061 (Therapeutic or Preventive Procedure, ex.: Cesarean section; Dermabrasion; Family psychotherapy)
- Drugs
    - T195 (Antibiotic, ex.: Antibiotics; bactericide; Thienamycins)
    - T200 (Clinical Drug, ex.: Ranitidine 300 MG Oral Tablet [Zantac]; Aspirin 300 MG Delayed Release Oral)
    - T121 (Pharmacologic Substance, ex.: Antiemetics; Cardiovascular Agents; Alka-Seltzer)

Both QuickUMLS and BERT were integrated into Spacy[3] pipeline over its preprocessing.

As a postprocessing, some stopwords that could not contribute to relevant clinical trials retrieving were removed, like, e.g. a single word "disease" marked as a Condition. If any abbreviations were marked as entities, they were replaced with full words using the list collected from Wikipedia[4].

Negations (expressions like "no" etc.) were also extracted by BERT and linked to the corresponding entities by word order and dependencies parsing from Spacy[5].

As for topics, there were also parsed age (using simple regular expressions) and gender (using patterns written for Spacy's EntityRuler[6]). There was no need to parse them from clinical trials since these fields were included into the XML schema of the trials suggested for the task.

3 Preliminary selection

Since running named entities extraction on all clinical trials suggested for the task would be computationally inefficient, a preliminary selection was made. The trials were put into a PostgreSQL database. Named entities recognition was run on the topics.

It seems that diseases are crucial for judging whether a patient is eligible for some trial. So a decision was made to perform the preliminary selection by conditions. Some patient ETH contain many conditions of different importance. The expressions having endings like "disease", "pathy", "itis" etc. were considered to be diseases (using a simple regular expression). Then they were sorted by their frequencies in Wikipedia[7]. (It seems that keywords for rare diseases and exact diagnoses are more valuable than more common ones).

The database was queried on age, gender, presence of the condition expression in inclusion criteria and absence in exclusion criteria until the number of trials reached some threshold. If it occurred already with the first query, procedures and drugs were also used to make the selection more strict. If no trials were found

[3] https://spacy.io/

[4] https://en.wikipedia.org/wiki/List_of_medical_abbreviations:_A

[5] https://spacy.io/usage/linguistic-features#dependency-parse

[6] https://spacy.io/api/entityruler

[7] https://github.com/IlyaSemenov/wikipedia-word-frequency

by all the conditions (or there were no conditions in the patient description, i.e. the person was healthy) the search was performed by procedures and drugs.

Then named entities recognition was run on the trials retrieved. The entities found in inclusion criteria linked to negations were considered to be exclusions (and the same for the opposite). Then the trials were again filtered by the exclusions.

4 Ranking

4.1 Using word frequencies

The results were left as they were after sorting by keywords frequencies in Wikipedia, without any further ranking. This approach corresponds to RUN1FREQS in the results submitted.

4.2 Using Word2Vec embeddings

BioWordVec [6] word embeddings model was used to calculate the cosine distance between topic conditions and trial conditions. For entities containing more than one word, embeddings of their words were averaged. Then cosine distances between each pair of query and trial entities were summarized. The trials were ranked according to these sums (the less the sum, the more relevant the trial is). This approach corresponds to RUN0 in the results submitted.

4.3 Using sentence-transformers embeddings

"paraphrase-TinyBERT-L6-v2"[8] sentence embeddings model [4] was used to produce embeddings for all the entities in text without directly averaging or summarizing embeddings for separate words. The entities were joined by commas. The trials extracted by a particular keyword at a preliminary selection step were ranked according to the cosine distances between an embedding of conditions in a query and in a trial. This approach corresponds to RUN3SENTS in the results submitted.

5 Results

	Runs		
	RUN1FREQS	RUN0	RUN3SENTS
ndcg_cut_5	0.3395	0.2567	**0.3757**
ndcg_cut_10	0.3024	0.2472	**0.3335**

Table 1: NDCGs over all topics

[8] https://huggingface.co/sentence-transformers/paraphrase-TinyBERT-L6-v2

	Runs		
	RUN1FREQS	RUN0	RUN3SENTS
P_10	0.2067	0.18	**0.22**

Table 2: Precision at 10 over all topics - eligible only

	Runs		
	RUN1FREQS	RUN0	RUN3SENTS
recip_rank	0.4036	0.3238	**0.4649**

Table 3: Reciprocal Rank over all topics - eligible only

Term frequencies approach (RUN1FREQS) can be considered to be a baseline. Averaged Word2Vec embeddings were found to be ineffective. Ranking by sentence-transformers embeddings showed the best result. Low precision at 10 numbers demonstrate the problem of too many exclusion criteria not taken into account.

5.1 Error Analysis

In this section the examples of the trials that got zero relevance will be given.

None of our models managed to successfully handle patient 4. There were a lot of conditions mentioned in her description, while the main one was pericardial effusion. Our models suggested trials on diseases that the patient had in the anamnesis but was not suffering from currently e.g. meningitis. It looks like pericardial effusion could not be caught with regular expression responsible for finding diseases that were used already on the preliminary selection step.

An opposite case was patient 69 who had no severe illnesses. She was suggested trials based on her procedures (e.g. vaccination). It turned out that some trials related to vaccination require pregnant patients, however, our models could not handle the logic that "menopausal" should be considered an exlusion criteria for trials with pregnant patients even if not explicitly mentioned in exclusion criteria.

6 Conclusion

Although the approach of selecting and ranking clinical trials based on named entities recognition can produce reasonable results, it has several limitations.

The main challenges are choosing named entities that characterize the current patient's condition best and handling exclusions not only by exact string matching, but also semantically.

Bibliography

[1] Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)

[2] Kury, F., Butler, A., Yuan, C., Fu, L.h., Sun, Y., Liu, H., Sim, I., Carini, S., Weng, C.: Chia, a large annotated corpus of clinical trial eligibility criteria. Scientific data **7**(1), 1–11 (2020)

[3] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[4] Reimers, N., Gurevych, I.: 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), `https://arxiv.org/abs/1908.10084`

[5] Soldaini, L., Goharian, N.: Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR workshop, SIGIR. pp. 1–4 (2016)

[6] Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z.: Biowordvec, improving biomedical word embeddings with subword information and mesh. Scientific data **6**(1), 1–9 (2019)