Using Eligibility Criteria to Rank Clinical Trials

Anonymous ACL submission

Abstract

Ranking clinical trials, patient descriptions pairs is a practical task for clinicians seeking to match patients with eligible trials. This task is introduced in the TREC Precision Medicine (PM) 2021 track, and shares some similarities to other query, document ranking tasks, like MS Marco Passage retrieval, and the TREC PM 2020 task, but also some important differences. Mainly, the length of the query[1] and document texts, and the document structure, which we show has an implied logical semantics that can be captured to improve ranking model performance. We evaluate a number of ways of representing both topics and documents using recall@100, recall@1000, and MRR (mean reciprocal rank), per similar approaches in the literature. We also fine-tune sciBERT on a portion of the data to evaluate a dense representation's ability to capture semantic similarity, as done in related work. Finally, we experiment with combinations of these scoring representations as features, with a look toward future applications.

1 Credits

This work serves to fill the requirement for a Masters in Computer Science through *Oregon Health and Science University*. Great thanks and admiration to all of the instructors who helped guide my degree process and research!

2 Introduction

The data used for these experiments comes from a set of clinical trials and paired relevance judgments supplied by Koopman & Zuccon (2016), hereafter referred to as the KZ dataset. Examples in the KZ dataset contain a set of patient descriptions and clinical trials documents, taken from a snapshot of the clinicaltrials.gov database in 2016. The patient descriptions are 5-10 sentences in length and taken from ER admission notes. Clinical trials are available in XML format, with a variety of more or less relevant fields, requiring some preprocessing, including sentence segmentation, to retrieve the desired sections of raw text. For this we developed a python package for open-source use for those interested in continuing or doing related work, described in the appendix section. For the experiments in this paper, only the 'brief summary', 'max_age', 'min_age', 'gender', and especially 'eligibility/criteria/textblock' fields are used, similar to how a human might parse the documents. An example patient description topic and clinical trials document annotated as relevant are shown below:

> **topic:** A 58-year-old African-American woman presents to the ER with episodic pressing/burning anterior chest pain that began two days earlier for the first time in her life. The pain started while she was walking, radiates to the back, and is accompanied by nausea, diaphoresis and mild dyspnea, but is not increased on inspiration. The latest episode of pain ended half an hour prior to her arrival. She is known to have hypertension and obesity. She denies smoking, diabetes, hypercholesterolemia, or a family history of heart disease. She currently takes no medications. Physical examination is normal. The EKG shows nonspecific changes.
>
> **exclude criteria**: Previous experience with a cardiac rehabilitation program. Patients with depression, uncontrolled diabetes and other significant comorbidities that may interfere with effective IHD management. Those patients, who in the mind of the attending

[1] 'query' and 'topic' are used interchangeably throughout this paper.

physician, are unsuitable for participation. Those unable to provide informed consent. Pregnant women. High-risk patients for safety considerations (future studies will high-risk patients.,

include criteria: Men and women admitted for an IHD event (acute coronary syndrome or revascularization procedure) who are at low or moderate risk.91 Regular Internet access (home, work or other environment). Over 18 years of age. Permission of the attending physician. Able to read, write and understand English without difficulty. No physical limitations to regular activity.

Given a patient description (topic) and a clinical trials (document) as seen above, determining relevance is a task usually done by medical experts. Some of the criteria can be very complex, implying multiple qualifers and subconditions. (see Appendix for examples). Sometimes the inclusion and exclusion structure is very loose, even intermixed. Because the number of irrelevant documents greatly exceed the number of relevant documents for a given query, the examples in the dataset, as in many ranking tasks, are taken from a pooled set of results from various other ranking processes. There are 3626 valid pairs like this one in the dataset. Some pairs were filtered out before calculating similarity for age, gender mismatch, or having no specified criteria. It is noted that others may not have done this and perhaps this contributed to a lower score than some others may find on this dataset, but this was ¡ 5% of examples, and wouldn't have contributed greatly in any case.

3 Related Work

Similar recent tasks include the MSMarco Passage Ranking task (Microsoft, 2021; Han, et al. (2020)) and the TREC PM 2019 task (TREC, 2019). In the passage ranking task, developed by Microsoft, the dataset consists of about 500k real Bing search queries and their most relevant results. At the time of this paper, success has been found with using BERT transformers to encode both query and representation and using similarity metrics (Lin 2021; Han, 2020), in addition to sparse representations, and often as a way of pre-ranking to pass a small subset to another model that has been trained to correctly classify the example pairs (Lin, et al. 2021). While in the CT task both query and document are on generally larger bodies of text than in the MS-MARCO task, another major difference between this task and the MSMarco task lies in large part to the difference in the document structure. The Clinical Trials (CT) eligibility criteria structure is not a single body of text that can be matched against, syntactically or semantically, because inclusion and exclusion sections imply different relationships between the query and topic than similarity alone. Results for this task were evaluated using MRR for ranking and recall for dataset reducing capability as a first-pass.

DK Dataset Description	
topics	60
docs	3626
2 rel pairs	2764
1 rel pairs	685
0 rel pairs	421
no elig criteria	16
no include	24
no exclude	180
doc age max range (yrs)	0.03-130
doc age min range (yrs)	0.08-80
no age req	2120
no gender req	3181
male req	366
female req	78
mean doc crit len	177
mean doc inc crit len	74
mean doc exc crit len	103
mean topic len	78

In the TREC PM 2019 task the queries are highly structured sets of cancer type, gene information, and demographics (not full sentences with various types of information that may or may not be relevant), and the documents are abstracts from PUBMED articles, once again, lacking an implied logical structure akin to the eligibility criteria field.

4 Methods

Typical methods for scoring topic and document similarity, using only the textual representations of the examples, include universal mapping, term-expansion, removal of stopwords and frequently occurring words, among others. Each of these methods were attempted with varying results, depending on the metric used. The best results of these methods came from the entity alias expansion. These were improved upon by splitting the inclusion and exclusion text and applying similarity *bm25* to each of these separately, then downweighting the exclusion empirically in a grid search. All of the methods are done on split portions of include

and exclude criteria separately, and only include text was evaluated for similarity if not specified otherwise. Similarity metrics were built-in methods to the pyserini library which was used to index the documents and provide a ranking by similarity score. *bm25* parameters were fixed at 0.9 and 0.4 per the findings of (Lin, et al., 2021). Also, importantly, methods in all cases were applied to both topic and document representations. Examples of all representations are in the appendix.

4.1 Universal Mapping

Many libraries exist for mapping terms to universal symbols, in the hopes of capturing a more generic representation, for various domains. In the biomedical domain, one commonly used representation is the UMLS (Universal Medical Language System) (NIH, 2021). These terms are then used to relate fields having differing syntax. e.g. neurofibromatosis ⇒ neurofibromatosis Type 2, benign neoplasm of cranial nerves ⇒ neurofibromatosis Type 2. This method, as well as sentence segmentation, was implemented using spaCy's models that have been trained on biomedical text (scispaCy).

4.2 Entity Alias Expansion

Along with UMLS terms, the spaCy processing pipeline produced raw text related to the particular UMLS token representation. The top 2 of alternate forms were taken (after some minimal experimentation with taking 1 or 3, achieving worse results), injected into the original text, with an effort to preserve sentence structure, though admittedly, the semantics become very unnatural when the sentences contain sections of redundant synonyms. Despite this unnaturality, this method performed the best of the spare representation methods tried.

4.3 Stopwords and Top Words

Experiments with removal of stopwords and N=100 most common words were attempted and applied before and after other methods, but failed to improve performance, similar to results found by these authors (Qiao, et al. 2019), hypothesized to be related to the ability of the model to ignore less useful tokens.

4.4 Moving Negated Statements

As mentioned, the logical implications of the criteria statements imply an AND function over the inclusion statements, and a NOT OR over the exclusion statements. While we understand this may not be as strict as this description supposes (for instance, allowing for some exclusion statements to have True values and keeping the document, topic pair relevant due to other, more weighted features), we observed that negated statements like the one shown in fig. 2 (figure out how to add image..) violate this strict formalisation. We then experiment (to no avail), with identifying the negated phrases as entities with negation labels (provided by spaCy implementation of the negex algorithm), move them as separate criteria to the opposing field (inclusion → exclusion, exclusion → inclusion).

4.5 Exclusion Weighting

Exlusion criteria were indexed and ranked for similarity in the same way as the inclusion criteria matching described in all of the methods here, and consistently produced the opposite results - that is, MRR on exclusion criteria was less than MRR for the exclusion criteria in all experiments. (see table for some examples)

Theoretically this can be explained by the similarity of a topic with an exclusion statement being a contribute to the *irrelevance* of a topic, document pair. We hypothesized that the weight of this exclusion criteria might be less or more than the weight of the inclusion criteria for a qiven topic, when combined in a linear expression, and did a grid search on $W \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5\}$,

and found $W = 0.75$ to be the best fit, for weighting the exclude similarity score output from the indexing, when combining with the inclusion similarity score for that pair, as shown in table 1, and formally in the equation below. Ultimately this should be a learned parameter in a downstream, final ranking, model.

$$similarity(topic, doc) = bm25(topic, doc.include) - W * bm25(topic, doc.exclude)$$

4.6 sciBERT fine-tuning

Though it is common to use pre-trained embedding representations for computing similarity scores, it has been shown that fine-tuning on the task dataset can greatly improve results (Lin et al., 2021). SciBERT is a transformer model similar to BERT (same learning tasks), but trained on a very different dataset (scibert, 2021), containing much more scientific data than the general BERT-base-uncased model, largely trained on Newscorp and Wikipedia

articles. But notably, not clinical trials data. Fine-tuning on a 90% training split of the data with a simplified classification task achieved marginal results ($\tilde{7}0\%$ f1), but did improve the models ability to construct embeddings that were better at matching with relevant docs than irrelevant ones, based on some preliminary tests. Loss from the classification model was weighted by class frequency before being backpropagated in order to compensate for class imbalance inherent in the data and to the task, and *f1* was used to calculate the loss.

Input examples passed to the sciBERT model consisted of concatenated [CLS] + topic + [SEP] + inc_criteria + [SEP] + exc_criteria, each of which was truncated or padded to achieve a standard input dimension of 512. (Output embedding dimension is 768). Once trained (after 5 epochs, with Adam optimization, lr = 2e-1, weight decay = 0.01), the model embeddings were obtained by taking the average of the last hidden layer outputs, as found to be the best method for this kind of similarity calculation by (Qiao, et al. 2019)

We tried a couple different approaches with the scoring of the embeddings. The first way took each of the criteria (inclusion, exclusion) separately, then we tried weighting as with the sparse experiments, but having a W=-1 gave the best results. We then took this combined score and further combined it with the normalized score from the sparse indexing results, using the best representations, the alias expansion. The results were kind of surprising, in that MRR was worse but recall of 1 (partially relevant) labelled pairs was 0.98!

5 Results

Table 1 shows MRR and precision at cutoffs of 100 and 1000. MRR is calculated simply as $1/rank(doc)$, where the best possible score is 1/1, the first relevant document being at rank 1. This is in contrast to precision@cutoff, which only values catching relevant documents in a net of a certain size, regardless of their position in the net. The first metric, MRR, rewards an algorithm for return a relevant result sooner. The second for obtaining many relevant results. Clearly, if only maximizing for precision of relevant documents, we could take all of the docs and assume they are relevant and catch all of them in our net, but then this is not a useful algorithm, as our set of possible documents is the same size as we started. By combining, taking the mean of, the 2 metrics, we can account for both goals. As we show, the goals are not always aligned for a given method.

The best results came from using the a combination of sciBERT fine-tuned embedding similarity and the sparse alias expansion similarity, with criteria weighting. This is consistent with the improved results of (ref) using BERT representations over purely *bm25* similarity, but note that we have a much smaller dataset, even if very differently structured, as described previously.

6 Discussion

This is a difficult, real-world task that is not represented well in the machine-learning literature to-date. Beyond simple text similarity this task requires a logical inference over the data pair, specific to the structure of include, exclude criteria. It could be framed as an NLI task in this respect, but the size of the text, and once again, the bi-valent structure of the documents, makes model typically trained for NLI sentence pairs less useful. Ongoing work could pursue several avenues, including: passing the useful features to a LeTOR supervised model like a Gradient Boosting Machine, to learn the weights to apply to the features to maximize relevance, other, possibly symbolic, graphical ways of representing medical relationships not captured by the UMLS used here, pre-training on the entire dataset (not just fine-tuning), and building up the dataset more, possibly using summarization to produce new examples. Finally, it is well understood that generally ranking models that only compare at the query, document level are limited when compared to models that are trained on examples of pairs of documents (one pos, one neg example), or lists of documents even, to learn a relative score instead of an absolute score. This presents a large computational burden, and it's not clear that there aren't ways to reduce the size of the text the model needs to learn, such as summarization, or attention at the sentence (criteria) level perhaps.

7 References

Lin, J., Ma, X., Lin. S-C, Yang, J-H, Pradeep, R. Nogueira, R. (2021). Pyserini: An Easy-to-Use Python Toolkit to Support Replicable IR Research with Sparse and Dense Representations.

Lin, J., Nogueira, R, Yates, A. (2021). Pretrained Transformers for Text Ranking: BERT and Beyond, retrieved 12/14/2021 from:

Method	MRR@10	MRR@100	$\text{recall}_{1}@100$	$\text{recall}_{2}@100$	$\text{recall}_{1}@1000$	$\text{recall}_{2}@1000$
concat all	0.19	0.20	0.24	0.30	0.59	0.65
remove stops	0.19	0.20	0.22	0.29	0.51	0.62
UMLS mapping	0.25	0.26	0.21	0.30	0.48	0.51
move negations	0.19	0.20	0.22	0.29	0.51	0.62
alias expansion	0.29	0.30	0.25	0.31	0.57	0.59
exclusion alone	0.11	0.12	0.13	0.17	0.37	0.39
crit weighting	0.30	0.30	0.21	0.18	0.54	0.18
sciBERT ft	0.65	0.65	0.56	0.60	0.57	0.61
sciBERT ft + cw	0.56	0.56	0.99	0.26	1.0	0.26

Table 1: Results for performance of various methods of representing doc, topics for ranking. All values rounded to hundredths place. sciBERT ft is actually a combination of equal weighted similarities calculations between the topic representation and both the exclude and include as separate representations.

https://arxiv.org/pdf/2010.06467.pdf

Koopman, B. Zuccon, G. (2016). A Test Collection for Matching Patients to Clinical Trials. *SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, July 2016 Pages 669–672, https://doi.org/10.1145/2911451.2914672

Microsoft MS Marco Dataset (2021). Description and data available at: https://microsoft.github.io/MSMARCO-Passage-Ranking/

National Institutes of Health (2021) UMLS, retrieved 12/14/2021 from https://www.nlm.nih.gov/research/umls/index.html

Qiao, Y., Liu, Zhe., Xiong, C., Liu, Zhi. (2004). Understanding the Behaviors of BERT in Ranking. Retrieved 12/14/2021 from: https://arxiv.org/pdf/1904.07531.pdf

scispacy (2021) retrieved from: https://allenai.github.io/scispacy/

Shuguang, H., Xuanhui, W., Bendersky, M. Najork, M. (2021). LEARNING-TO-RANK WITH BERT IN TF-RANKING. retrieved 12/14/2021 from: https://arxiv.org/pdf/2004.08476.pdf

TREC Precision Medicine 2019 track link: http://www.trec-cds.org/2019.html

A Appendices

B Supplemental Material

B.1 sparse representation methods

Original Criteria:

Inclusion Criteria:
- Patients with HF or IHD who are not currently taking the study medications of interest (ACE inhibitors/angiotensin receptor blockers for HF or statins for IHD) and whose primary care physicians are part of the study population

Exclusion Criteria:
- Patients who are unable or unwilling to give informed consent,
- previously taken the study medications according to dispensing records
- allergy or intolerance to study medications
- residents of long-term care facilities
- unable to confirm a diagnosis of either HF or IHD
- primary care physician has already contributed 5 patients to the study

need way to break out of multicolumn for supplemental.

B.1.1 concatenating all relevant fields

B.1.2 removing stopwords

B.1.3 UMLS value mapping (CUIs)

B.1.4 moving negations to opposing criteria field

B.1.5 alias expansion