Trial2Vec: Zero-Shot Clinical Trial Document Similarity Search using Self-Supervision

Anonymous ACL submission

Abstract

Clinical trials are essential for drug development but are extremely expensive and time-consuming to conduct. It is beneficial to study similar historical trials when designing a clinical trial. However, lengthy trial documents and lack of labeled data make trial similarity search difficult. We propose a *zero-shot* clinical trial retrieval method, called Trial2Vec, which learns through self-supervision without the need for annotating similar clinical trials. Specifically, the *meta-structure* of trial documents (e.g., title, eligibility criteria, target disease) along with clinical knowledge (e.g., UMLS knowledge base [1]) are leveraged to automatically generate contrastive samples. Besides, Trial2Vec encodes trial documents considering meta-structure thus producing compact embeddings aggregating multi-aspect information from the whole document. We show that our method yields medically interpretable embeddings by visualization and it gets 15% average improvement over the best baselines on precision/recall for trial retrieval, which is evaluated on our labeled 1600 trial pairs. In addition, we prove the pretrained embeddings benefit the downstream trial outcome prediction task over 240k trials.

1 Introduction

Clinical trials are essential for developing new medical interventions (Friedman et al., 2015). Many considerations come into the design of a clinical trial, including study population, target disease, outcome, drug candidates, trial sites, and eligibility criteria, as in Table 1. It is often beneficial to learn from related clinical trials from the past to design an optimal trial protocol. However, accurate similarity search based on the lengthy trial documents is still in dire need.

Self-supervision based pretraining has delivered promising performances for many NLP and CV tasks with fine-tuning (Devlin et al., 2019; Liu et al., 2019; He et al., 2021; Bao et al., 2021). Nevertheless, we find there was few work on *zero-shot document retrieval* as most address document retrieval in a supervised fashion (Humeau et al., 2019; Khattab and Zaharia, 2020; Guu et al., 2020; Karpukhin et al., 2020; Lin et al., 2020; Luan et al., 2021; Wang et al., 2021; Hofstätter et al., 2020; Li et al., 2020; Zhan et al., 2021; Hofstätter et al., 2021b,a; Jiang et al., 2022) or improve document pre-training for further supervision (Beltagy et al., 2020; Zaheer et al., 2020; Ainslie et al., 2020; Zhang et al., 2021).

Recently, a burgeoning body of research (Gao et al., 2021; Wu et al., 2021; Wang et al., 2022) proposes to execute self-supervised learning to train semantic-meaningful *sentence* embeddings free of labels. However, there are still challenges to apply them for document similarity search:

- **Lengthy documents.** These zero-shot BERT re-

Table 1: An example of the meta-structure of clinical trial document drawn from *ClinicalTrials.gov*.

Title	Effects of Electroacupuncture With Different Frequencies for Major Depressive Disorder
Description	Two groups of subjects will be included 55 subjects in electroacupuncture with 2Hz group...
Eligibility Criteria	1. Inclusion Criteria: 1.1. Patients suffering from MDD in accordance with the diagnostic criteria; 1.2. Hamilton Depression Scale score is between 21 and 35 (mild to moderate MDD);... 2. Exclusion Criteria: 2.1 Patients with bipolar disorder; 2.2 Patients with schizophrenia or other mental disorders; ...
Outcome Measures	1. Change in anxiety and depression severity measure by Self-rating depression scale 2. Change in the severity of depression measure by Hamilton depression scale ..
Disease	Major Depressive Disorder
Intervention	electroacupuncture
...	...

[1] https://www.nlm.nih.gov/research/umls/index.html

trieval methods all work on short sentences (usually below 10 words) similarity search while trial documents are often above 1k words. Simply encoding lengthy trials by truncating and averaging embeddings of all remaining tokens inevitable leads to poor retrieval quality.

- **Inefficient contrastive supervision.** These unsupervised methods take simple instance discriminative contrastive learning (CL) within batch, e.g., SimCSE (Gao et al., 2021) takes one sentence into the encoder twice to get the positive pairs and all other sentences as the negative. This paradigm has low supervision efficiency to require a large batch size, large data, and long training time, which is infeasible for learning from long trial documents.

In this work, we propose Clinical **Trial TO Vec**tors, `Trial2Vec`, a zero-shot trial document similarity search using self-supervision. We design a trial encoding framework considering the meta-structure to rid the risk that semantic meaning vanishes due to the uniform average of token embeddings. Meanwhile, the meta-structure is utilized to generate contrastive samples for efficient supervision. Medical knowledge is introduced to further enhance the negative sampling for CL. Our main contributions are:

- We are the first to study the trial-to-trial retrieval task by proposing a label-free SSL model which is able to encode long trials into semantic meaningful embeddings without labels.
- We propose a data-efficient CL method on medical knowledge and trial meta-structure, which is promising to be extended to further zero-shot structured document retrieval.
- We demonstrate the superiority of `Trial2Vec` on a trial relevance dataset of 1600 trials annoated by domain experts. Also, we show `Trial2Vec` can assist better downstream trial outcome prediction on a dataset of 240k trials.

2 Related works

2.1 Text & document retrieval

General texts. Early information retrieval methods depend on manual engineering (Robertson and Zaragoza, 2009; Yang et al., 2017). By contrast, dense retrieval methods based on distributional word representations, e.g., Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), Doc2Vec (Le and Mikolov, 2014), etc., become popular crediting to their superior performance. The advent of deep models, especially the contexualized encoders like BERT (Devlin et al., 2019), encourages an explosion of neural retrieval methods (Van Gysel et al., 2016; Zamani et al., 2018; Guo et al., 2016; Dehghani et al., 2017; Onal et al., 2018; Reimers and Gurevych, 2019; Chang et al., 2019; Nogueira and Cho, 2019; Chen et al., 2021; Lin et al., 2020; Xiong et al., 2020; Karpukhin et al., 2020; Yates et al., 2021). However, most of them are based on supervised training on sentence pairs from general texts, e.g., SNLI (Bowman et al., 2015). When label is expensive to acquire, as in the clinical trial case, we need zero-shot learning models. Although, there arose some works to perform post-processing on pretrained BERT embeddings to improve their retrieval quality (Li et al., 2020; Su et al., 2021), their performances are far from optimal without specific training.

Clinical trials. Traditional clinical trial query search systems (Tasneem et al., 2012; Tsatsaronis et al., 2012; Jiang and Weng, 2014; Park et al., 2020) are established on protocol databases. Contrast to dense retrieval, these methods rely on entity matching with rules thus not flexible enough. Recent works (Roy et al., 2019; Rybinski et al., 2020, 2021) propose supervised neural ranking for clinical trial query search. However, all of them work on matching trial titles or relevant segments with an input user query. While `Trial2Vec` can also assist query search, it is the first to encode complete trial documents for the trial-level similarity search.

2.2 Text contrastive learning

Contrastive learning is a heated discussed topic recently in NLP and CV (Chen et al., 2020a,b; Chen and He, 2021; Carlsson et al., 2020; Zhang et al., 2020; Wu et al., 2020; Yan et al., 2021; Gao et al., 2021). CL is one main topic under the SSL domain. It sheds light on reaching comparable performance as supervised learning free of manual annotations. While CL has been applied to enhance downstream NLP applications like text classification (Li et al., 2021; Zhang et al., 2022), a few (Wang et al., 2020; Zhang et al., 2020; Yan et al., 2021; Yang et al., 2021) are able to do zero-shot retrieval. Nonetheless, all focus on enhancing *sentence* embeddings by manipulating text only therefore are suboptimal when facing lengthy documents. By contrast, `Trial2Vec` uses the doc-

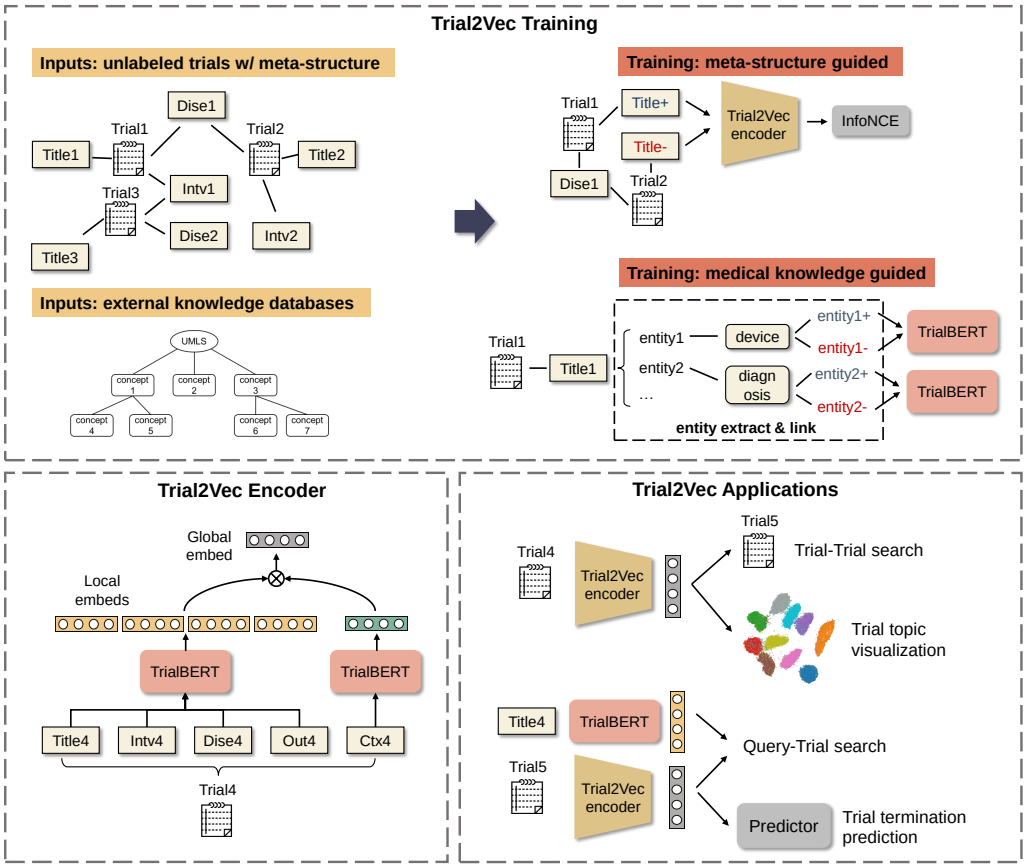

Figure 1: Overview of the proposed `Trial2Vec` framework. **Top left**: the training strategy that accounts for unlabeled input trial documents with meta-structure along with an external medical knowledge database, e.g., UMLS. **Top right**: The contrastive supervision splits into meta-structure and knowledge guided, respectively. **Bottom left:** our method hierarchically encodes trials into local and global embeddings on the trial meta-structure. **Bottom right:** The encoded trial-level embeddings can be used to trial search, query trial search and downstream tasks.

ument meta-structure with domain knowledge to obtain and facilitate *document* embeddings.

3 Method

In this section, we present the details of `Trial2Vec`. The main idea is to jointly learn the global and local representations from trial documents considering their meta-structure. Specifically, observed in Table 1, trial document consists of multiple sections while the *key attributes* (e.g., title, disease, intervention, etc.) occupy a small portion of the whole document. This motivates us to design a hierarchical encoding and the corresponding contrastive learning framework. The overview is illustrated in Fig. 1. Our method generates *local attribute embeddings* using the `TrialBERT` backbone separately, then aggregating local embeddings with a learnable attention module to obtain the *global trial embeddings* that emphasize significant attributes. We present the pretraining of backbone encoder in §3.1; then we describe the hierarchical encoding process based on the backbone encoder in §3.2; the hierarchical constrastive learning methods considering meta-structure and medical knowledge are elucidated in §3.3; at last, we elicit the applications of the proposed framework in §3.4.

Table 2: List of text corpora used for continual pretraining of `TrialBERT`.

Corpus	Number of words
ClinicalTrials.gov	240M
Medical Encyclopedia	3M
Wikipedia Articles	11M

3.1 Backbone encoder: `TrialBERT`

We leverage the BERT architecture as the backbone encoder in the framework. In detail, we use the WordPiece tokenizer together with the BioBERT (Lee et al., 2020) pretrained weights as the start

point. We continue the pretraining with Masked Language Modeling (MLM) loss on three trial-related data sources: ClinicalTrial.gov [2], Medical Encyclopedia [3], and Wikipedia Articles [4], see Table 2, to get `TrialBERT`. ClinicalTrials.gov is a database that contains around 400k clinical trials conducted in 220 countries. Medical Encyclopedia has 4K high-quality articles introducing terminologies in medicine. We also retrieve relevant Wikipedia articles corresponding to the 4k terminologies of Medical Encyclopedia.

3.2 Global and local embeddings by `Trial2Vec`

`TrialBERT` embeddings pretrained with MLM on clinical corpora still hold weak semantic meaning. Meanwhile, previous sentence embedding BERTs all take an average pooling over token embeddings, which causes the semantic meaning vanishing when applied to lengthy clinical trials. Therefore, we propose `Trial2Vec` architecture that exploits the *global* and *local* embeddings for trial based on its meta-structure.

We split the attributes of a trial into two distinct sets: *key attributes* and *contexts*. The first component includes the trial title, intervention, condition, and main measurement, which are sufficient to retrieve a pool of coarsely relevant trial candidates; the second includes descriptions, eligibility criteria, references, etc., which differentiate trials targeting similar diseases or interventions because they provide the multi-facet details regarding disease phases, study designs, targeted populations, etc. According to this design, local embeddings $\{\mathbf{v}_{att}\}_{l=1}^{L} \in \mathbb{R}^{L\times D}$ are produced separately on each key attribute. On the other hand, a context embedding is obtained by encoding the context texts $\mathbf{v}_{ctx} \in \mathbb{R}^{D}$. Note that the above encoding is all conducted by the same encoder.

We further refine the local embeddings by context embeddings and aggregate them to yield the global trial embedding $\mathbf{v}_g \in \mathbb{R}^D$. The refinement is performed by multi-head attention, as

$$\mathbf{v}_g = \texttt{MultiHeadAttn}(\mathbf{v}_{ctx}, \{\mathbf{v}_l\}_l^L, \mathbf{W}), \quad (1)$$

which relocates the attention over key attributes to enhance discrminative power of the yielded global embedding.

[2] https://clinicaltrials.gov/

[3] https://medlineplus.gov/encyclopedia.html

[4] https://www.wikipedia.org/

3.3 Hierarchical contrastive learning

For data-efficient contrastive learning, we utilize the meta-structure & medical knowledge for contrasting local and global embeddings hierarchically.

Global contrastive loss. The first objective is to maximize the semantic in trial embeddings for similarity search. Instead of doing in-batch instance-wise contrastive loss like SimCSE, we propose to sample informative negative pairs by exploiting the trial meta-structure. As shown by Fig. 1, some trials may be linked by a common attribute like disease or intervention. Denote a trial consisting of several attributes by

$$\mathbf{x} = \{x^{\text{title}}, x^{\text{intv}}, x^{\text{dise}}, x^{\text{out}}, x^{\text{ctx}}\}, \quad (2)$$

we can build an informative negative sample by replacing its title with a trial which also targets for disease x^{dise} by

$$\mathbf{x}^- = \{x^{\text{title}-}, x^{\text{intv}}, x^{\text{dise}}, x^{\text{out}}, x^{\text{ctx}}\}. \quad (3)$$

Meanwhile, we apply a random attribute dropout towards $\mathbf{x}$ to formulate a positive sample as

$$\mathbf{x}^+ = \{x^{\text{title}}, x^{\text{dise}}, x^{\text{out}}, x^{\text{ctx}}\}. \quad (4)$$

InfoNCE loss is utilized in a batch of B trials as

$$\mathcal{L}_g = -\sum_{i=1}^{B} \log \frac{\exp(\psi(\mathbf{v}_{gi}, \mathbf{v}_{gi}^+))}{\sum_{v_{gi}^- \in \mathcal{V}_i^-} \exp(\psi(\mathbf{v}_{gi}, \mathbf{v}_{gi}^-))}, \quad (5)$$

where the negative sample set $\mathcal{V}_i^- = \{\mathbf{v}_{gi}^-\} \cup \{\mathbf{v}_{gj}\}_{j\neq i}$; $\psi(\cdot,\cdot)$ measures the cosine similarity between two vectors. The global contrastive loss here encourages the model to capture the attribute of interest by discriminating the subtle differences of input trial attributes, which prevent the semantic meanings from vanishing due to the average pooling over all trial texts.

Local contrastive loss. In addition to the global trial embeddings, we put supervision on local embeddings to inject medical knowledge into the model. Unlike general texts, two medical texts can be overlapped word-wise dramatically but still describe two distinct things[5], which is challenging for similarity computing. To strengthen `TrialBERT` discriminative power for medical texts, we extract key medical entities in each text as [6]

$$E(x^{att}) = \{e_1, e_2, e_3, e_4\}, \quad (6)$$

[5] For instance, replacing *Olaparib* in "*A Phase I, Open-Label, 2 Part Multicentre Study to Assess the Safety and Efficacy of Olaparib*" with another intervention like *Vitamin D* renders a total different study topic.

[6] Done by SciSpacy https://scispacy.apps.allenai.org/.

then a positive sample is built by mapping one entity e_1 to its canonical name or a similar entity under the same parental conception $\hat{e}_1$ defined by UMLS as

$$E(x^{att+}) = \{\hat{e}_1, e_2, e_3, e_4\}. \quad (7)$$

Similarly, negative sample is built by deletion or replacing one entity with another dissimilar one. InfoNCE loss is therefore used by

$$\mathcal{L}_l = -\sum_{i=1}^{B} \log \frac{\psi(\mathbf{v}_{li}, \mathbf{v}_{li}^{+})}{\sum_{\mathcal{V}_{li}^{-}} \exp(\psi(\mathbf{v}_{li}, \mathbf{v}_{li}^{-}))}. \quad (8)$$

We at last jointly optimize the global and contrastive losses as

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_l. \quad (9)$$

3.4 Application of global & local embeddings

The hierarchical contrastive learning offers extraordinary flexibility of `Trial2Vec` for various downstream tasks in *zero-shot* learning. At first, the global trial embeddings $\mathbf{v}_g$ can be directly used for similarity search by comparing trial pair-wise cosine similarities. The computed trial embeddings can also help identify and discover research topics when we apply visualization techniques. On the other hand, we can also execute query search using partial attributes crediting to the contrastive learning between local and global embeddings. When we need do trial-level predictive tasks, e.g., trial termination prediction, a classifier can be attached to the pretrained global trial embeddings and learned; the backbone `TrialBERT` is also capable of offering short medical sentence retrieval because of local contrastive learning.

4 Experiments

In this section, we conduct five types of experiments to answer the following research questions:

- **Exp 1 & 2.** How does `Trial2Vec` perform in complete and partial retrieval scenarios?
- **Exp 3.** How do the proposed SSL tasks / embedding dimension contribute to the performance?
- **Exp 4.** Is the trial embedding space interpretable and aligned with medical ontology?
- **Exp 5.** How useful do well-trained `Trial2Vec` contribute to downstream tasks, e.g., trial outcome prediction, after fine-tuned?
- **Exp 6.** Qualitative analysis of the retrieval results and what are the differences of `Trial2Vec` and baselines?

Table 3: Statistics of trial status in *ClinicalTrials.gov* database where we conclude *Approved* & *Completed* as completion; *Suspended*, *Terminated*, and *Withdrawn* as the termination for trial outcome prediction.

Approved	Completed	Suspended	Terminated	Withdrawn
174	210,237	1,658	22,208	10,439
Available	Enrolling	Unavailable	Not recruiting	Recruiting
237	3,662	45,128	18,171	60,362
Completion	Termination	Summary	Others	
210,411	34,305	244,716	127,560	

4.1 Dataset & Setup

Trial Similarity Search. We created a labeled trial dataset to evaluate the retrieval performance where paired trials are labeled as relevant or not. We keep 311,485 interventional trials from the total 399,046 trials. We uniformly sample 160 trials as the query trials. To overcome the sparsity of relevance, we take advantage of TF-IDF (Salton et al., 1983) to retrieve ranked top-10 trials as the candidate to be labeled, resulting in 1,600 labeled pairs of clinical trials. Unlike general documents, the clinical trial document contains many medical terms and formulations. We recruited clinical informatics researchers, and each is assigned 400 pairs to label as relevant or not using label $\{1, 0\}$. To keep labeling processes in line, we specify the minimum annotation guide for judging relevance: (1) same disease; or (2) same intervention and similar diseases (e.g., cancer on distinct body parts). We use precision@k (prec@k) and recall@k (rec@k) to evaluate and report retrieval performances, where

$$prec@k = \frac{\text{\# of relevant trials in the top k results}}{k}, \quad (10)$$

$$rec@k = \frac{\text{\# of relevant trials in the top k results}}{\text{\# of relevant trials in all candidate trials}}. \quad (11)$$

Trial termination prediction. We can take the pretrained `Trial2Vec` embeddings for predicting the trial outcomes, i.e., if the trial will be terminated or not. We add one additional fully-connected layer on the tail of `Trial2Vec`. The targeted outcomes are in the status section of clinical trials, described by Table 3. We formulate the outcome prediction as a binary classification problem to predict the *Completion* or *Termination* of trials where we get 210,411 and 34,305 trials as positive and negative labeled, respectively. We take 70% of all as the training set and 20% as the test set; the remaining 10% is used as the validation set for tuning and

Table 4: Precision/Recall of the retrieval models on the labeled test set. Values in parenthesis show 95% confidence interval. Best values are in bold.

Method	prec@1	prec@2	prec@5	rec@1	rec@2	rec@5
TF-IDF	0.5132(0.063)	0.4386(0.045)	0.3828(0.057)	0.1871(0.038)	0.3172(0.026)	0.6147(0.044)
Word2Vec	0.7492(0.071)	0.6476(0.044)	0.4712(0.033)	0.3008(0.054)	0.4929(0.042)	0.7939(0.041)
TrialBERT	0.7264(0.050)	0.6219(0.060)	0.4324(0.027)	0.3257(0.051)	0.4896(0.054)	0.7611(0.041)
BERT-Whitening	0.7476(0.094)	0.6630(0.045)	0.4525(0.029)	0.3672(0.045)	0.5832(0.042)	0.8355(0.021)
BERT-SimCSE	0.6788(0.039)	0.5995(0.035)	0.4714(0.021)	0.2824(0.034)	0.4566(0.035)	0.8098(0.025)
`Trial2Vec`	**0.8740(0.026)**	**0.7524(0.049)**	**0.5027(0.055)**	**0.4053(0.066)**	**0.6449(0.060)**	**0.8769(0.030)**

early stopping. We utilize three metrics for evaluation: accuracy (ACC), area under the Receiver Operating Characteristic (ROC-AUC), and area under Precision-Recall curve (PR-AUC).

4.2 Baselines & Implementations

We take the following baselines for retrieval: TF-IDF (Salton et al., 1983; Salton and Buckley, 1988), Word2Vec (Mikolov et al., 2013), BERT-Whitening (Huang et al., 2021; Su et al., 2021), and BERT-SimCSE (Gao et al., 2021). Details of these methods can be seen in Appendix A.

We keep all methods' embedding dimensions at 768. We start from a BERT-base model to continue pre-training on clinical domain corpora, yielding our `TrialBERT`, which supports as the backbone for BERT-Whitening and BERT-SimCSE for fair comparison. We take 5 epochs with batch size 100 and the learning rate 5e-5. In the second SSL training phase, AdamW optimizer with a learning rate of 2e-5, batch size of 50, and weight decay of 1e-4 is used. Experiments were done with 6 RTX 2080 Ti GPUs.

4.3 Exp 1. Complete Trial Similarity Search

Since labels are unavailable in the training phase, we only chose unsupervised/self-supervised baselines. Results are shown by Table 4. `Trial2Vec` outperforms all baselines with a great margin. It has around 15% improvement on each metrics than the best baselines on average. For baselines, all except for TF-IDF have similar performance. When k is small, the precision gap between `Trial2Vec` and baselines is large; when k is large, all methods encounter precision reduction. That is because the pool of candidate trials are 10 but the number of positive pairs for each are often less than 5, which limits the maximum of the numerator of $prec@k$ in Eq. (10). Likewise, `Trial2Vec` also shows stronger performance in $rec@k$ because it is discounted by the maximum number of positive pairs.

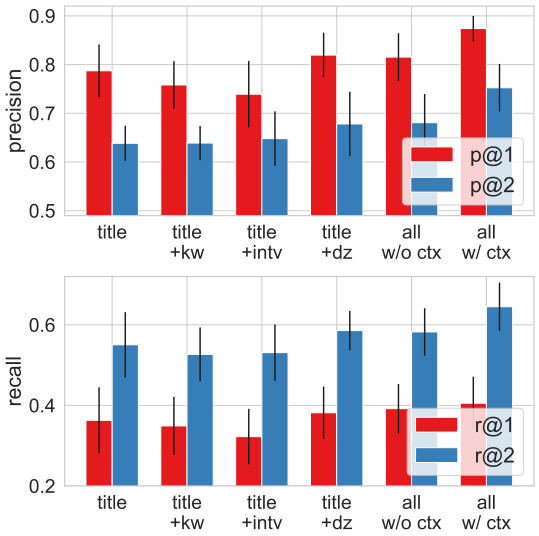

Figure 2: Performance of `Trial2Vec` on the partial retrieval scenarios. We use a different part of the trial as queries to retrieve similar trials, including keyword *kw*, intervention *intv*, disease *dz*, context *ctx*. Error bars indicate the 95% confidence interval of results.

Interestingly, the state-of-the-art sentence BERTs, e.g., BERT-whitening and BERT-simCSE, have limited improvement over original BERT and even Word2Vec. Unlike general documents, clinical trials may be overlapped in much content but still be irrelevant if the key entities are different. This special characteristic causes the assumption of *a document with similar passage is relevant* (Craswell et al., 2020) used in general document retrieval but invalidated in clinical trial retrieval. Without well-designed SSL, it is hard for these methods to learn these subtle differences. Moreover, clinical trial documents are often much longer than the general documents in those open datasets. There are 622.4 words per trial on average, while the general STS benchmark has below 15 words per sample, e.g., STS-12: 10.8, STS-13: 8.8, STS-14: 9.1, etc (Cer et al., 2017). We also observed the simple negative sampling strategy of SimCSE

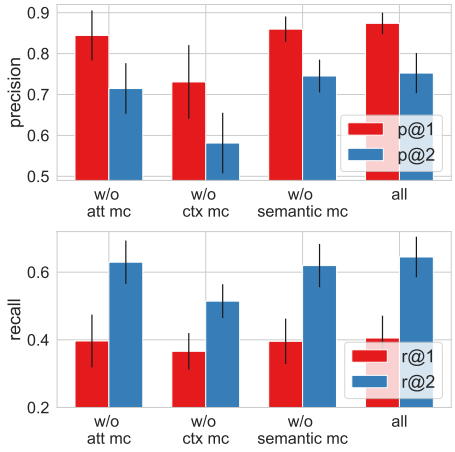

Figure 3: Ablation study on the contribution of each Task to the final result. *att*, *mc*, *ctx* are short for attribute, matching, context, respectively. *all* indicate the full `Trial2Vec` that all tasks are used.

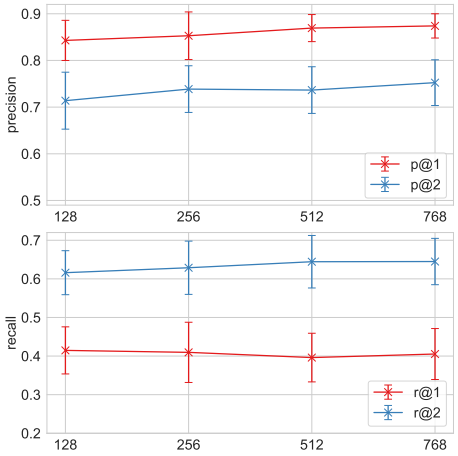

Figure 4: Analysis of the influence of embedding dimensions on retrieval quality by `Trial2Vec`: embedding dim in 128, 256, 512, 768. Error bars show the 95% confidence interval.

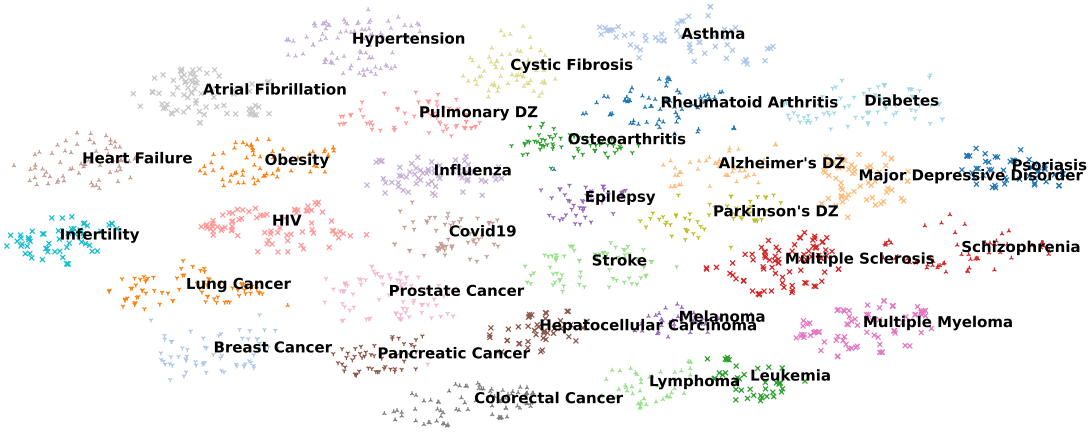

Figure 5: 2D visualization of the trial-level embeddings obtained by `Trial2Vec` (dimension reduced by t-SNE). It can be seen trials are automatically classified into clusters by topic (diseases) in the embedding space. For example, a series of tumor-related trials (e.g., Breast and Pancreatic Cancers) are on the bottom of the embedding space.

is insufficient to learn effective long document embeddings. In comparison, `Trial2Vec` leverages the meta-structure of clinical trials to focus on the most informative attributes, with additional context-based refinement, producing embeddings superior in semantic representation.

Table 5: Trial outcome prediction performances of baselines and `Trial2Vec`, after fine-tuned.

Method	ACC	ROC-AUC	PR-AUC
TF-IDF	0.8571(0.002)	0.7194(0.004)	0.2960(0.008)
Word2Vec	0.8574(0.002)	0.7189(0.005)	0.2906(0.007)
TrialBERT	0.8559(0.002)	0.7277(0.006)	0.3109(0.006)
`Trial2Vec`	**0.8622(0.002)**	**0.7332(0.004)**	**0.3137(0.007)**

4.4 Exp 2. Partial Query Trial Retrieval

We further investigate the partial trial retrieval scenario where users intend to find similar trials with short and incomplete descriptions, e.g., partial attributes. Results are illustrated by Fig 2. We start by measuring how well `Trial2Vec` only utilizes the title for trial retrieval. It is witnessed that using title is sufficient to yield comparable performance as the best baseline for complete retrieval shown in Table 4. Nonetheless, we identify that concatenating keywords or intervention with the title reduces performance. Combining title and disease yields similar performance as involving all attributes. This phenomenon signifies that the disease plays a vital role in trial similarity and is always recommended to be involved in query trial retrieval.

Table 6: Case studies comparing the retrieval performance of the Trial2Vec with baseline models. Due to the space limits, only title and NCT ID of trials are given.

Query Trial	TF-IDF	TrialBERT	Trial2Vec
[NCT02972294] HiFIT Study : Hip Fracture: Iron and Tranexamic Acid (HiFIT)	[NCT01221389] Study Using Plasma for Patients Requiring Emergency Surgery (SUPPRES)	[NCT04744181] Patient Blood Management In CARdiac sUrgical patientS (ICARUS)	[NCT01535781] Study of the Effect of Tranexamic Acid Administered to Patients With Hip Fractures. Can Blood Loss be Reduced?
[NCT01590342] Diclofenac for Submassive PE (AINEP-1)	[NCT04006145] A Phase 2 Study of Elobixibat in Adults With NAFLD or NASH	[NCT04156854] Intravascular Volume Expansion to Neuroendocrine-Renal Function Profiles in Chronic Heart Failure	[NCT00247052] Non Steroidal Anti Inflammatory Treatment for Post Operative Pericardial Effusion

4.5 Exp 3. Ablation Studies

We conducted ablation studies to measure how SSL tasks and embedding dimensions contribute to final results. Results are shown by Fig. 3, where we remove one Task for each setting and reevaluate. Here, *att mc* and *ctx mc* corresponds to the global contrastive loss by negative sampling on key attributes and contexts, respectively; *semantic mc* indicates the local contrastive loss. We observe that ctx mc is very important. Without it, only attributes of trials are included in the training and inference of Trial2Vec, thus resulting in a significant performance drop. However, even only using a small segment of trials (the attributes), Trial2Vec still reaches similar performance as BERT-SimCSE that receives the whole trial document as inputs. This demonstrates the importance of picking high-quality negative samples during the CL process. Similarly, we observe other two tasks also improve the retrieval quality.

Fig. 4 illustrates the retrieval performance on different embedding dimensions. We identify that reducing embedding dimension does not affect the performance of Trial2Vec much, i.e., one can choose a small embedding dimension (e.g., 128) without suffering much performance degradation while saving lots of storage and computational resources.

4.6 Exp 4. Embedding Space Visualization

Fig. 5 plots the 2D visualization of the embedding space of Trial2Vec using t-SNE (Van der Maaten and Hinton, 2008) where around 2k trials uniformly sampled from 300k trials. The tag texts illustrate the target diseases of trials with different colors. We observe that these trials embeddings show interpretable clusters corresponding to target disease categories. More discussions about this visualization can be referred to Appendix B.

4.7 Exp 5. Trial Termination Prediction

Results are illustrated by Table 5. Compared with the shallow models, BERT-based methods gain better performance, which credits the deep architecture of transformers with stronger learning capability. Trial2Vec takes a hierarchical encoding for trial documents on meta-structure thus better revealing the trial characteristics, which plays a central role in predicting its potention outcomes.

4.8 Exp 6. Case Study

We perform a qualitataive analysis of similarity search results and two baselines. Results are shown in Table 6. These two case studies show that TF-IDF and BERT models all tend to put attention on frequent words in query trials, e.g., *blood* and *iron* in case study 1; and *heart failure* in case study 2. This bias comes from the average pooing taken onto all token embeddings. The top-1 relevant clinical trial retrieved by Trial2Vec, on the other hand, provides a more similar trial thanks to the hierarchical encoding and specific local and global contrastive learning. We add more explanations regarding these cases in Appendix C.

4.9 Conclusion

This paper investigated utilizing BERT with self-supervision for encoding trial into dense embeddings for similarity search. Experiments show our method can succeed in zero-shot trial search under various settings. The embeddings are also useful for trial downstream predictive tasks. The qualitative analysis, including embedding space visualization and case studies, further verifies that Trial2Vec gets a medically meaningful understanding of clinical trials.

References

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: Encoding long and structured inputs in transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284.

Hangbo Bao, Li Dong, and Furu Wei. 2021. BEiT: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Wei-Cheng Chang, X Yu Felix, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2019. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33:22243–22255.

Xiaoyang Chen, Kai Hui, Ben He, Xianpei Han, Le Sun, and Zheng Ye. 2021. Co-BERT: A context-aware bert retrieval model incorporating local and query-specific context. *arXiv preprint arXiv:2104.08523*.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 55–65.

Lawrence M. Friedman, Curt D. Furberg, David L. DeMets, David M. Reboussin, and Christopher B. Granger. 2015. *Fundamentals of Clinical Trials*. Springer, New York, NY.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *ACM International on Conference on Information and Knowledge Management*, pages 55–64.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021a. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.

Sebastian Hofstätter, Bhaskar Mitra, Hamed Zamani, Nick Craswell, and Allan Hanbury. 2021b. Intra-document cascading: Learning to select passages for neural document ranking. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1349–1358.

Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and Allan Hanbury. 2020. Local self-attention over long text for efficient document retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2021–2024.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan

Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. *arXiv preprint arXiv:2104.01767*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Silis Y Jiang and Chunhua Weng. 2014. Cross-system evaluation of clinical trial search engines. *AMIA Summits on Translational Science Proceedings*, 2014:223.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. PromptBERT: Improving bert sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196. PMLR.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *EMNLP*.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv preprint arXiv:2010.11386*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Kezban Dilek Onal, Ye Zhang, Ismail Sengor Altingovde, Md Mustafizur Rahman, Pinar Karagoz, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, et al. 2018. Neural information retrieval: At the end of the early years. *Information Retrieval Journal*, 21(2):111–182.

Junseok Park, Seongkuk Park, Kwangmin Kim, Woochang Hwang, Sunyong Yoo, Gwan-su Yi, and Doheon Lee. 2020. An interactive retrieval system for clinical trial studies with context-dependent protocol elements. *PloS one*, 15(9):e0238290.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Soumyadeep Roy, Koustav Rudra, Nikhil Agrawal, Shamik Sural, and Niloy Ganguly. 2019. Towards an aspect-based ranking model for clinical trial search. In *International Conference on Computational Data and Social Networks*, pages 209–222. Springer.

Maciej Rybinski, Sarvnaz Karimi, and Aleney Khoo. 2021. Science2Cure: A clinical trial search prototype. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2620–2624.

Maciej Rybinski, Jerry Xu, and Sarvnaz Karimi. 2020. Clinical trial search: Using biomedical language understanding models for re-ranking. *Journal of Biomedical Informatics*, 109:103530.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Gerard Salton, Edward A Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Asba Tasneem, Laura Aberle, Hari Ananth, Swati Chakraborty, Karen Chiswell, Brian J McCourt, and Ricardo Pietrobon. 2012. The database for aggregate analysis of clinicaltrials. gov (aact) and subsequent regrouping by clinical specialty. *PloS one*, 7(3):e33677.

George Tsatsaronis, Konstantinos Mourtzoukos, Vassiliki Andronikou, Tassos Tagaris, Iraklis Varlamis, Michael Schroeder, Theodora Varvarigou, Dimitris Koutsouris, and Nikolaos Matskanis. 2012. PONTE: a context-aware approach for automated clinical trial protocol design. In *proceedings of the 6th International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases in conjunction with VLDB*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2016. Learning latent vector spaces for product search. In *ACM International on Conference on Information and Knowledge Management*, pages 165–174.

Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. SNCSE: Contrastive learning for unsupervised sentence embedding with soft negative samples. *arXiv preprint arXiv:2201.05979*.

Shuohang Wang, Yuwei Fang, Siqi Sun, Zhe Gan, Yu Cheng, Jingjing Liu, and Jing Jiang. 2020. Cross-thought for sentence encoder pre-training. In *Conference on Empirical Methods in Natural Language Processing*, pages 412–421.

Zifeng Wang, Rui Wen, Xi Chen, Shilei Cao, Shao-Lun Huang, Buyue Qian, and Yefeng Zheng. 2021. Online disease diagnosis with inductive heterogeneous graph convolutional networks. In *Proceedings of the Web Conference*, pages 3349–3358.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021. ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. *arXiv preprint arXiv:2109.04380*.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1256.

Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2021. Universal sentence representation learning with conditional masked language model. In *Conference on Empirical Methods in Natural Language Processing*, pages 6216–6228.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *ACM International Conference on Web Search and Data Mining*, pages 1154–1156.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *ACM International Conference on Information and Knowledge Management*, pages 497–506.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512.

Hang Zhang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Poolingformer: Long document modeling with pooling attention. In *International Conference on Machine Learning*, pages 12437–12446. PMLR.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Conference on Empirical Methods in Natural Language Processing*, pages 1601–1610.

Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang, and Jiawei Han. 2022. Metadata-induced contrastive learning for zero-shot multi-label text classification. *arXiv preprint arXiv:2202.05932*.

A Baselines for clinical trial similarity search

- TF-IDF (Salton et al., 1983; Salton and Buckley, 1988). It is short for term frequency–inverse document frequency that has been widely used for information retrieval systems for decades. One can use TF-IDF for document retrieval by concatenating scores of all words in this document then computing cosine distance between document vectors.

- Word2Vec (Mikolov et al., 2013). It is a classic dense retrieval method by building distributed word representations by self-supervised learning methods (CBOW). We take an average pooling of word representations in a document for retrieval by cosine distance.

- `TrialBERT`. We take an average pooling over all token embeddings at the last layer of it for similarity computation.

- BERT-Whitening (Huang et al., 2021; Su et al., 2021). This is an unsupervised post-processing method that uses anisotropic BERT embeddings (Ethayarajh, 2019; Li et al., 2020) to improve semantic search. We take the average of last and first layer of its BERT embeddings following Su et al. (2021).

- BERT-SimCSE (Gao et al., 2021). It is a contrastive sentence representation learning method stemming from InfoNCE loss. It simply takes other samples in batch as negative samples.

B Embedding space visualization

From Fig. 5, trial embeddings are clearly clustered into topics with self-supervised learning, which provides a great help for topic mining and discovery for the existing clinical trials. For instance, we can find that cancers that happen on different body parts are near to each other on the bottom of the embedding space (Prostate Cancer, Breast Cancer, Pancreatic Cancer, Colorectal Cancer, etc.). Also, the diseases which are related to brain function, e.g., Alzheimer's Disease, Parkinson's Disease, Major Depressive Disorder, etc. Other examples include Covid19, Influenza, Pulmonary Disease, etc.

The reason is that we explicitly utilize the knowledge from attributes of trials for negative sample building, which endows the embedding space the ability to discriminate trials' similarity. These similar trials can also have similar characteristics like having similar recruiting criteria or targeting similar outcome measures, which are captured by `Trial2Vec` by refining the embeddings of attributes by detailed descriptions. Based on this observation, we can infer that such medically meaningful trial embeddings would be beneficial to downstream tasks on clinical trials, e.g., trial outcome prediction.

C Case Study

For the first case, the query trial is [NCT02972294], which studies using Tranexamic acid and Iron Isomaltoside to reduce the occurrence of Anemia and blood transfusion in hip fracture cases. We show the top-1 retrieved by three methods on the right. Trial found by TF-IDF studies the efficiency of plasma in patients with Hemorrhagic shock; BioBERT finds a trial about patients undergoing heart surgery who have Anaemia to test if a correction of iron reduces red blood cell transfusion requirements. `Trial2Vec` finds a trial that studies Tranexamic acid effect in blood loss in hip fracture operations. `Trial2Vec` result is highly relevant to the query trial as it has the identical drug on blood loss of the same type of operation.

In the second example, the query trial tries to investigate the benefits of Diclofenac for Normotensive patients with acute symptomatic Pulmonary Embolism and Right Ventricular Dysfunction. TF-IDF finds an irrelevant study on the efficacy and safety of Elobixibat for adults with NAFLD or NASH. `TrialBERT` also retrieves an irrelevant study on Intravascular Volume Expansion to Neuroendocrine-Renal Function Profiles in Chronic Heart Failure. On the other hand, `Trial2Vec` digs out a trial that studies the same type of drug with a similar purpose as the target's: evaluating the efficiency of NSAID (Diclofenac) to the evolution of postoperative (cardiac surgery) pericardial effusion.

D Potential limitations & risks of this work

The empirical evaluation of this method is mainly done on the clinical trial documents drawn from *ClinicalTrials.gov* which were fully written in English. It might be the best fit when this method is applied to documents in other languages. Although

we have tried our best to collect trial relevance datasets, it is still possible that the datasets used for evaluation are not able to cover all cases.

The proposed framework encodes trial documents into compact embeddings for search. It encounters failure cases some time as wrong trials are retrieved. It should be used with discretion when applied to clinical trial research or by individual volunteers who intend to look for trials research. Retrieved results in practice should be used under the supervision with professional clinicians.