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

Zifeng Wang¹ and Jimeng Sun¹,²

¹Department of Computer Science, University of Illinois Urbana-Champaign
²Carle Illinois College of Medicine, University of Illinois Urbana-Champaign
{zifengw2,jimeng}@illinois.edu

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.²

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 (Wang et al., 2022b). 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; Wang et al., 2022c). 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; Karpukin 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., 2021a,b; 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., 2022a) 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:

| 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                                                                      |
| ...                    | ...

¹https://www.nlm.nih.gov/research/umls/index.html
²Software is available at https://github.com/RyanWangZf/Trial2Vec.

Table 1: An example of the meta-structure of clinical trial document drawn from ClinicalTrials.gov.
• **Lengthy documents.** These zero-shot BERT retrieval 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 inevitably 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 Vectors, 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 annotated 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 contextualized 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; Wang et al., 2020b; Wang and Sun, 2022). 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., 2020a; 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 document 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 contrastive 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.

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 

3https://clinicaltrials.gov/
Encyclopedia⁴, and Wikipedia Articles⁵, see Table 6, 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 trials; 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 \( \{\mathbf{v}_{att}\}^l_{i=1} \in \mathbb{R}^{L \times D} \) 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

\[
\mathbf{v}_g = \text{MultiHeadAttn}(\mathbf{v}_{ctx}, \{\mathbf{v}_i\}^L_l, \mathbf{W}),
\]

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

### 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

\[
x = \{x^{\text{title}}, x^{\text{intv}}, x^{\text{dise}}, x^{\text{out}}, x^{\text{ctx}}\},
\]

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

\[
x^- = \{x^{\text{title}_-}, x^{\text{intv}}, x^{\text{dise}}, x^{\text{out}}, x^{\text{ctx}}\}.
\]

Meanwhile, we apply a random attribute dropout towards \( x \) to formulate a positive sample as

\[
x^+ = \{x^{\text{title}}, x^{\text{dise}}, x^{\text{out}}, x^{\text{ctx}}\}.
\]

InfoNCE loss is utilized in a batch of \( B \) trials as

\[
\mathcal{L}_g = -\sum_{i=1}^{B} \log \frac{\exp(\psi(\mathbf{v}_{gi}, \mathbf{v}_{gj}^+))}{\sum_{g\in V_i^-} \exp(\psi(\mathbf{v}_{gi}, \mathbf{v}_{gj}^-))},
\]

where the negative sample set \( V_i^- = \{\mathbf{v}_{gj}\}_{j\neq i} \cup \{\mathbf{v}_{gj}\}_{j\neq i}; \psi(\cdot, \cdot) \) 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 things⁶, which is challenging for similarity computing. To strengthen TrialBERT discriminative power for medical texts, we extract key medical entities in each text as

\[
E(x^{att}) = \{e_1, e_2, e_3, e_4\},
\]

⁶For 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.

⁷Done 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
\[
E(x^{att+}) = \{\hat{e}_1, e_2, e_3, e_4\}. \tag{7}
\]
Similarly, negative sample is built by deletion or replacing one entity with another dissimilar one. InfoNCE loss is therefore used by
\[
L_l = -\sum_{i=1}^{B} \log \frac{\psi(v_{li}, v_{li}^+)}{\sum_{v_{li}^-} \exp(\psi(v_{li}, v_{li}^-))}. \tag{8}
\]
We at last jointly optimize the global and contrastive losses as
\[
L = L_g + L_l. \tag{9}
\]

### 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 \( v_g \) 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 2: 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.**

| Status     | Approved | Completed | Suspended | Terminated | Withdrawn |
|------------|----------|-----------|-----------|------------|-----------|
| Count      | 210,237  | 1,658     | 22,208    | 18,171     | 60,362    |
| Percent    | 210,411  | 34,305    | 45,128    | 127,560    | 63,810    |

### 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), recall@k (rec@k), and nDCG@5 to evaluate and report performances.

\[
\text{prec@k} = \frac{\# \text{ of relevant trials in the top k results}}{k}, \tag{10}
\]
\[
\text{rec@k} = \frac{\# \text{ of relevant trials in the top k results}}{\# \text{ of relevant trials in all candidate trials}}. \tag{11}
\]

**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 2. 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 3: Precision/Recall and nDCG 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   | nDCG@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) | 0.5480(0.034) |
| BM25         | 0.7015(0.044) | 0.5640(0.041) | 0.4246(0.032) | 0.3538(0.038) | 0.4841(0.050) | 0.7666(0.031) | 0.7312(0.033) |
| 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) | 0.7712(0.032) |
| BERT         | 0.7264(0.050) | 0.6219(0.060) | 0.4324(0.027) | 0.3257(0.051) | 0.4896(0.054) | 0.7611(0.041) | 0.7370(0.047) |
| BERTWhiten   | 0.7476(0.094) | 0.6630(0.045) | 0.4525(0.029) | 0.3672(0.045) | 0.5832(0.042) | 0.8355(0.021) | 0.8129(0.024) |
| BERTSimCSE   | 0.6788(0.039) | 0.5995(0.035) | 0.4714(0.021) | 0.2824(0.034) | 0.4566(0.035) | 0.8098(0.025) | 0.7308(0.038) |
| MonoT5Med    | 0.6799(0.068) | 0.5810(0.061) | 0.4439(0.051) | 0.2904(0.032) | 0.4657(0.049) | 0.7570(0.037) | 0.7171(0.043) |
| Trial2Vec    | 0.8810(0.026) | 0.7912(0.049) | 0.5055(0.039) | 0.4216(0.046) | 0.6465(0.060) | 0.8919(0.030) | 0.8825(0.029) |

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), BM25 (Trotman et al., 2014), Word2Vec (Mikolov et al., 2013), BERT-Whitening (Huang et al., 2021; Su et al., 2021), BERT-SimCSE (Gao et al., 2021), and MonoT5 (Roberts et al., 2021; Pradeep et al., 2022). 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 3. 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 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.

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 that all tasks are used.

Most informative attributes, with additional context-based refinement, producing embeddings superior in semantic representation.

### 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. Combining title and concatenating keywords or intervention with the title reduces performance. Combining title and

| 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)** |

Table 4: Trial outcome prediction performances of baselines and Trial2Vec, after fine-tuned.
Table 5: 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 Study                                                                 | TrialBERT Study                                                              | Trial2Vec Study                                                                 |
|------------------------------------------------------------------------------|------------------------------------------------------------------------------|------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| [NCT03972394] HiFIT Study : Hip Fracture: Iron and Tranexamic Acid (HiFIT)  | Study Using Plasma for Patients Requiring Emergency Surgery (SUPPRES)         | Patient Blood Management In CARdiac sUrgi-cal patientS (ICARUS)               | 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)                         | Study of Elobixibat in Adults With NAFLD or NASH                             | Intravascular Volume Expansion to Neuroendocrine-Renal Function Profiles in Chronic Heart Failure |

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 4. 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 potential outcomes.

4.8 Exp 6. Case Study
We perform a qualitative analysis of similarity search results and two baselines. Results are shown in Table 5. 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 pooling 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.

5 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 vi-
visualization and case studies, further verifies that Trial2Vec gets a medically meaningful understanding of clinical trials.

**Acknowledgement**

This work was supported by NSF award SCH-2205289, SCH-2014438, IIS-1838042, NIH award R01 1R01NS107291-01.

**Limitations**

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.

**References**

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvikec, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghao, 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. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6894–6910.
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.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jiemeng Sun. 2022c. MedCLIP: Contrastive learning from unpaired medical images and text. In EMNLP.

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.
Table 6: List of text corpora used for continual pretraining of TrialBERT.

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

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.

- **BM25** (Trotman et al., 2014). A bag-of-words retrieval method commonly used in practice. We run it based on the rank-bm25 package \(^8\) with its default hyperparameters.

- **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. We use gensim \(^9\) to run this method.

- **BERT.** We take an average pooling over all token embeddings at the last layer of it for similarity computation. We take the TrialBERT pretrained on all the clinical trial documents.

- **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.

- **MonoT5-Med** (Pradeep et al., 2022). It was proposed in (Roberts et al., 2021) for matching patient descriptive texts and clinical trial documents via T5 model (Raffel et al., 2020) based on prompts. We use its version fine-tuned on Med Marco dataset (Koopman and Zuccon, 2016).

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.

\(^8\)https://pypi.org/project/rank-bm25/
\(^9\)https://radimrehurek.com/gensim/models/doc2vec.html
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 Elohexibat 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.