# Trial2Vec: Clinical Trial Similarity Search using Self-Supervised Siamese BERT model

Anonymous Author(s)

## ABSTRACT

Clinical trials are essential for drug development but extremely expensive and time-consuming to conduct. It is beneficial to study similar historical trials when designing a clinical trial. However, complex multi-aspect trial information and lack of labeled data make trial similarity search difficult. We propose a clinical trial retrieval method, called Trial2Vec. Unlike most BERT-based retrieval methods, Trial2Vec learns through self-supervision with the need for labeling a large number of clinical trial pairs. Trial2Vec produces compact embeddings for trials by considering the document structures of clinical trial reports. Moreover, introducing medical knowledge via self-supervision encourages Trial2Vec to model trial topics. Thus Trial2Vec yields medically interpretable embedding and superior performance for downstream clinical trial tasks such as trial outcome prediction with fine-tuning. We demonstrate that Trial2Vec produces 15% average improvement over the best baselines on the precision/recall, which is evaluated on our labeled 1,600 paired trial data. Additional qualitative experiments on embedding visualization and case studies verify that Trial2Vec gains medically interpretable representations of trials.

## 1 INTRODUCTION

Clinical trials are essential for developing new medical interventions [14]. Many considerations go into the design of a clinical trial, including study population, target disease, outcome, drug candidates, trial sites, and eligibility criteria. It is often beneficial to learn from related clinical trials from the past to design an optimal trial protocol. ClinicalTrials.gov[1] provides detailed multifaceted textual information about clinical trials from a wide range of medical interventions and disease conditions. Practitioners can search clinical trial records via keyword queries and simple faceted searches for disease conditions and intervention types. However, accurate similarity search based on complex and lengthy text descriptions is still lacking.

With the development of contextualized language models for broad natural language processing (NLP) tasks, e.g., BERT [12], dense retrieval methods become widely used in information retrieval practices [17, 19, 25, 26, 41]. The idea of dense retrieval is to encode documents into dense embedding vectors. A document can thus be ranked on the similarity of the encoded embeddings.

[1]https://clinicaltrials.gov/


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

**Table 1: An example of clinical trial record drawn from ClinicalTrials.gov. We only show the part of the whole record of this trial here.**

| | |
|---|---|
| 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 55 subjects in the electroacupuncture with 100Hz group. The clinical efficacy of electroacupuncture with different frequencies (2 Hz, 100 Hz) in the treatment of MDD will be observed by evaluation indicators, such as Self-rating depression scale and Hamilton depression scale. |
| Eligibility Criteria | 1. Inclusion Criteria:<br>1.1. Patients suffering from MDD in accordance with the diagnostic criteria;<br>1.2. Hamilton Depression Scale score is between 21 and 35 (mild to moderate MDD);<br>1.3. 18≤age≤60 years, both gender;...<br>2. Exclusion Criteria:<br>2.1 Patients with bipolar disorder;<br>2.2 Patients with schizophrenia or other mental disorders; ... |
| Outcome Measures | 1. Change in anxiety and depression severity measure by Self-rating depression scale<br>2. Change in the severity of depression measure by Hamilton depression scale .. |
| Disease | Major Depressive Disorder |
| Intervention | electroacupuncture |

However, there are still challenges applying BERT-based retrieval for clinical trials:

- **Weak semantic meaning of BERT embeddings**. The sentence/document embeddings from pretrained BERT perform poorly on similar retrieval tasks without fine-tuning [15, 24, 37]. Previous works propose fine-tuning BERT on large-scale paired query/document data, thus explicitly enhancing the BERT embeddings for semantic search [33]. Unfortunately, supervised fine-tuning cannot be applied directly due to the lack of labeled data from clinical trial retrieval tasks.
- **Multi-aspect information**. Unlike general document retrieval, clinical trials have rich structures, which combine tabular, numeric data and texts. As shown by Table 1 which is an example drawn from ClinicalTrials.gov, a trial contains unstructured fields, including plain texts such as summary and descriptions, semi-structured texts like eligibility criteria, and structured texts like outcomes and their measure. Thus, leveraging this complex

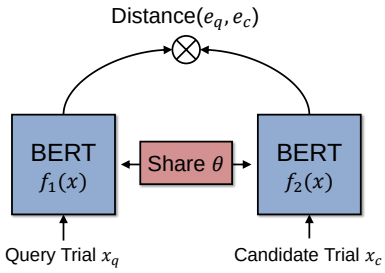


**Figure 1: Workflow of the Siamese structure in `Trial2Vec`. The query trial $x_q$ and the candidate ranking trial $x_c$ are processed by the dual BERT encoders, which share the same parameters. Then, the distance of encoded embeddings $e_q$ and $e_c$ are computed for ranking.**

structure and dealing with multi-modality data is a central challenge in trial representation and retrieval.

To tackle these challenges, we propose Clinical **Trial TO Vec**tors, `Trial2Vec`, a self-supervised learning (SSL) method using a Siamese BERT. This Siamese model can encode different trials into a common embedding space to enable an effective trial similarity search. We design effective SSL tasks that consider trials' structure and the medical domain knowledge. Our contributions are:

- We are the first to study the retrieval method for clinical trials by proposing `Trial2Vec`, which sheds light on expediting a wide range of clinical research applications by deep learning, e.g., trial outcome prediction.
- We execute BERT pre-training on large-scale medical and clinical related corpora. Then, we fine-tune it with the well-designed SSL tasks considering trial structure and clinical domain knowledge.
- To evaluate the retrieval performance of clinical trials, we collect a 1,600 trial relevance dataset with the assistance of domain experts. Our method yields 15% average improvement over the best baselines on precision/recall on this real-world dataset.

## 2 RELATED WORKS

### 2.1 Document Retrieval

Early information retrieval works mainly rely on BM25 [34, 45]. But these algorithms heavily depend on manual engineering, and is difficult to support downstream prediction tasks. In contrast, dense retrieval methods have become popular thanks to the success of machine learning. Pioneering works leverage dense distributional representation of documents, e.g., Word2Vec [27], Glove [31], and Doc2Vec [22], for computing text/passage similarity [10, 20, 21, 48]. Later, the advancement of deep learning catalyzed a series of neural information retrieval methods [11, 16, 30, 39, 49]. These DL-based retrieval methods are free of handcrafted features and take advantage of the powerful representation capability of neural networks, thus outperforming previous shallow methods.

Recent advance of contextualized pre-trained language models, e.g., BERT [12], delivers a burgeoning body of BERT based retrieval methods [4, 8, 19, 25, 29, 33, 43, 47]. However, they are based on supervised training on large-scale query-document pairs from general corpora, e.g., SNLI [1]. Post-processing techniques such as BERT-flow [24] and BERT-whitening [37] were proposed to improve the semantic representation of pre-trained BERT embeddings. Though free of labels, their performances are still far from satisfying in practice.

### 2.2 Self-supervised Learning & Contrastive Learning

Contrastive learning (CL) first raised great attention in the computer vision field [5–7]. Motivated by them, CL was introduced to NLP for textual representation learning [2, 15, 42, 44, 50]. These methods removed the requirement for training labels and still achieved promising performance comparable to supervised methods in various tasks. Moreover, CL can also provide valuable embeddings for downstream tasks. Most existing works are on open datasets like the STS benchmark [3], which contain general corpus and plain texts, thus not directly applicable to clinical trials. For example, SimCSE [15] takes the augmented view of the same sentence as the positive sample and all the other samples in batch as the negative samples for CL. However, SimCSE does not utilize any structure information of documents, thus weak in sampling informative negative samples, resulting in its need for a large batch size. As most texts in the open datasets are short sentences, e.g., STS-12 has only 10.8 words per sample [3], large batch CL is feasible. By contrast, the clinical trial document has a rich structure and has a length far longer than the limit of BERT inputs, e.g., the average length of trial document used in our experiments is 622.4, which renders these methods infeasible to train or leads to low accuracy.

## 3 METHOD

In this section, we first elaborate on the model architecture and input/output of the model in §3.1. Then, we introduce the two-stage training of `Trial2Vec`: pre-training (§3.2) and self-supervised training (§3.3).

As shown by the flowchart in Fig. 2, `Trial2Vec` encodes a trial in two parts: *key attributes* and *detailed descriptions*, accounting for the multi-modality of clinical trials. Attributes sketch up the whole trial on targeted disease, candidate interventions, and the proposed outcome measures. In most scenarios, using only key attributes are sufficient to retrieve a large pool of coarsely relevant trial candidates. In this sense, we propose two self-supervised tasks: *attribute matching* (§3.3.1) and *semantic matching* (§3.3.3), to enhance the semantic searching ability of attribute embeddings.

To further leverage the trial information to refine the embeddings, we leverage the detailed descriptions into `Trial2Vec` encoding process with a multi-head cross-attention module. This yields a context-aware embedding that considers the differences between trials with similar target diseases/interventions. These differences may be multi-facet, e.g., disease phases, study designs, targeted populations, which are reflected by detailed descriptions. We propose to do *context matching* (§3.3.2) to facilitate the context-aware embeddings.

### 3.1 `Trial2Vec` Inference

We will begin with the case of inference. For notations to be used please refer to Table 2. As mentioned above, `Trial2Vec` encodes

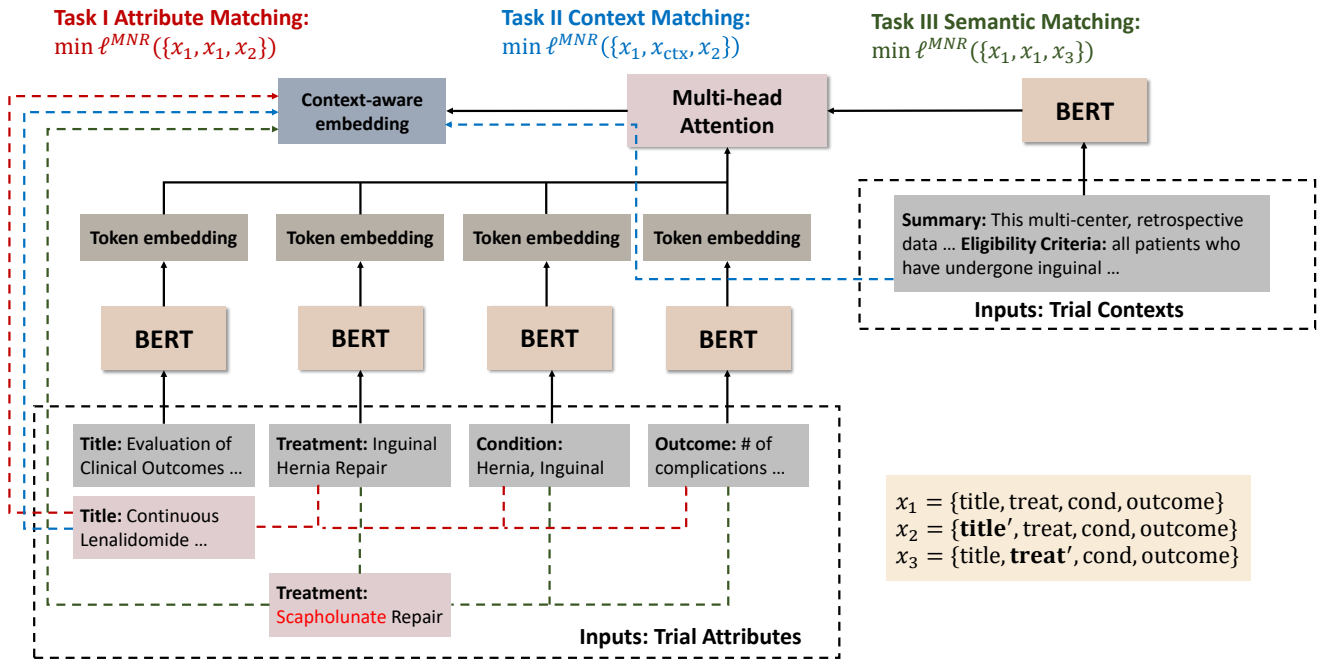


**Figure 2: The self-supervised learning framework of Trial2Vec: (1) Task I: attribute matching; (2) Task II: context matching; and (3) Task III: semantic matching. In Task I, we substitute one attribute of the trial to build negative sample for contrastive learning (CL) on the attribute embeddings. In Task II, we replace one attribute of the trial for CL on the context-attentive embeddings. In Task III, we replace or remove key entities in attributes to build negative samples for CL on the attribute embeddings.**

**Table 2: Notation table.**

| Symbol | Meaning |
|---|---|
| $x_i$ | Input clinical trial document |
| $x_{i,+}, x_{i,-}$ | Positive/Negative sample built based on $x_i$ |
| $x_i^a$ | One attribute (e.g. title) of a clinical trial document |
| $x_{i,j}^a$ | The $j$-th token in $x_i^a$ |
| $x_i^c$ | One context (e.g., eligibility criteria) of trial |
| $\mathcal{A}, C$ | Attribute and Context set of trial |
| $\boldsymbol{h}_{i,j}^a$ | Token embedding of $x_{i,j}^a$ |
| $\boldsymbol{h}_i^a$ | Aggregated embedding for $x_i^a$ |
| $\boldsymbol{h}_i^c$ | Aggregated embedding for $x_i^c$ |
| $\boldsymbol{h}_i^{\mathcal{A}}$ | Aggregated attribute embedding of $x_i^a$ where $a \in \mathcal{A}$ |
| $\boldsymbol{h}_i$ | Context-aware document embedding of $x_i$ |
| $\boldsymbol{h}_i'$ | $\boldsymbol{h}$ of $x_i$ with a different dropout mask in inference |

a trial in two distinct paths: key attributes and detailed descriptions. More specifically, we introduce four attributes for an input document $x_i$:

- Title ($x_i^{title}$): The title of the trial;
- Intervention ($x_i^{intv}$): The treatment of the trial, such as a drug or surgical procedure. For multiple interventions in one trial, we can concatenate them together to serve the encoder. It is the same for the other attributes;
- Condition ($x_i^{cond}$): The condition of interest, such as a disease like diabetes;
- Outcome ($x_i^{otc}$): The outcome of the clinical trial, such as blood pressure, body temperature, or tumor size.

Since we do not discriminate the context parts, we use $x_i^{ctx}$ to represent all of them. Three sections of trials are treated as the context:

- Summary: The brief introduction of background, motivation, and aim of the trial.
- Eligibility criteria: The inclusion/exclusion rules for recruiting individuals into the trial.
- References: The academic papers relevant to the trial.

For simplicity of discussion, we will use $\mathcal{A}$ to denote the set of attributes. Attributes $x_i^a$ are tokenized then fed into the BERT encoder to yield the token-level embeddings:

$$\boldsymbol{h}_{i,1}^a, \ldots, \boldsymbol{h}_{i,n_i^a}^a = \text{BERT}(x_{i,1}^a, \ldots, x_{i,n_i^a}^a) \text{ for } a \in \mathcal{A}, \tag{1}$$

where $n_i^a$ denotes the number of tokens in $x_i^a$; $x_{i,j}^a$ is the $j$-th token of attribute $x_i^a$ . To obtain the attribute-level embedding, one can take the average pooling within each attribute, resulting in $\boldsymbol{h}_i^a$ as

$$\boldsymbol{h}_i^a = \frac{1}{n_i^a} \sum_{j=1}^{n_i^a} \boldsymbol{h}_{i,j}^a. \tag{2}$$

When only partial trial information is known, e.g., we only have the title and disease information, we can take the aggregated embeddings of title and disease by average pooling

$$\boldsymbol{h}_i^{\mathcal{A}} = 1/|\mathcal{A}^{\dagger}| \sum \boldsymbol{h}_i^a \tag{3}$$

for retrieval, where $\mathcal{A}^{\dagger}$ is the known partial set of attributes.

On the other side, we obtain the context embedding by the average pooling of the encoded concatenated context as

$$\boldsymbol{h}_i^{ctx} = \frac{1}{n_i^c} \sum \texttt{BERT}\left(\texttt{Concat}(\{x_i^c\}_{c\in C})\right). \tag{4}$$

Here, $n_i^c$ is the number of tokens of all contexts; $x_i^c$ is one of the contexts; $C$ is the known context set.

Then, $\boldsymbol{h}_i^{ctx}$ is leveraged to refine the attribute embeddings through a multi-head attention module, as

$$\boldsymbol{h}_i = \texttt{MultiHead}\left(\boldsymbol{h}_i^{ctx}, \{\boldsymbol{h}_i^a\}_a, \{\boldsymbol{h}_i^a\}_a; K_h\right) = \texttt{Concat}\left(\{\tilde{\boldsymbol{h}}_{i,k}\}_{k=1}^{K_h}\right)\mathbf{W}^O. \tag{5}$$

Here, $K_h$ is the number of heads; $\mathbf{W}^O$ maps the concatenated heads embeddings $\tilde{\boldsymbol{h}}_{i,k}$ to the input size; and $\tilde{\boldsymbol{h}}_{i,k}$ are attentive embeddings obtained by

$$\tilde{\boldsymbol{h}}_{i,k} = \texttt{Attention}\left(\boldsymbol{h}_i^{ctx}, \{\boldsymbol{h}_i^a\}_a, \{\boldsymbol{h}_i^a\}_a\right) = \sum_{a\in\mathcal{A}} \alpha_i^a \mathbf{W}_k^V h_i^a, \tag{6}$$

where $\alpha_i^a$ is the attention weight generated by softmax scoring function computed by dot-product similarity between query $\mathbf{W}^Q \boldsymbol{h}_i^{ctx}$ and key $\mathbf{W}^K \boldsymbol{h}_i^a$; $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are three weight matrices. In a nutshell, $\boldsymbol{h}_i$ is the final embedding for the complete input clinical trial document $x_i$, and we could use it for downstream tasks.

## 3.2 Trial2Vec Pre-training

The training of Trial2Vec consists of two steps: (1) continual pretraining by masked language modeling and (2) fine-tuning by self-supervised learning. Step (1) adapts BERT to clinical domain and step (2) enriches the semantic meaning of trials into the dense embeddings.

The pre-training of our Trial2Vec begins at the checkpoint of BioBERT-base model [23] which is pretrained on large biomedical corpora including PubMed Abstracts and PMC Full-text articles using masked language modeling loss. We further enhance the pretraining on three clinical trial domain sources: ClinicalTrial.gov [2], Medical Encyclopedia [3], and Wikipedia Articles [4], see Table 3. 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 in Medical Encyclopedia.

During the pre-training, we use the WordPiece tokenizer which is used in both BERT and BioBERT [12, 23]. This standard tokenization enables the transfer of existing applications on BERT/BioBERT to Trial2Vec. It is also able to tokenize any new terms appearing in clinical trials.

[2]https://clinicaltrials.gov/
[3]https://medlineplus.gov/encyclopedia.html
[4]https://www.wikipedia.org/

**Table 3: List of text corpora used for continual pretraining of Trial2Vec.**

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

## 3.3 Trial2Vec Self-supervised Learning

The BERT encoder pre-trained with domain-specific corpora is helpful for downstream tasks. However, these embeddings still have weak semantic meaning, thus inferior for dense retrieval. Unlike the Task of question answering on the general domain as done in previous works [32, 40, 46], labeling the relevance of clinical trials is tremendously expensive because of the need for medical expertise. Therefore, we propose to train Trial2Vec for dense retrieval by SSL to avoid the requirement for labeled data.

In detail, we exploit the macro structure of clinical trial documents with two matching tasks: **Attribute Matching** and **Context Matching**. We also propose knowledge-driven contrastive learning based on Unified Medical Language System (UMLS) [5], namely **Semantic Matching**, to enhance the micro embeddings' semantic meaning in the clinical language.

### 3.3.1 *Task I: Attribute Matching.*

Denote the attributes of a trial document by

$$x_i^{\mathcal{A}} = \{x_i^{title}, x_i^{intv}, x_i^{cond}, x_i^{otc}\}, \tag{7}$$

we build the negative samples for it by replacing one arbitrary attribute from another trial $x_m$, $m \neq 1$, as

$$x_{i,-}^{\mathcal{A}} = \{x_m^{title}, x_i^{treat}, x_i^{cond}, x_i^{otc}\}. \tag{8}$$

Meanwhile, the other $N-1$ samples in the same batch $\{x_m\}_{m=1}^N \setminus \{x_i\}$ are also treated as negative samples, which renders the batch multiple negative ranking (MNR) loss

$$\ell_i^{AM} = \ell^{MNR}\left(x_i^{\mathcal{A}}, x_i^{\mathcal{A}}, \mathcal{X}^{\mathcal{A}}\right) = -\log \frac{\exp(\psi(\boldsymbol{h}_i^{\mathcal{A}}, \boldsymbol{h}_i^{\mathcal{A},\prime}))}{\sum_{x_m^{\mathcal{A}}\in\mathcal{X}} \exp(\psi(\boldsymbol{h}_i^{\mathcal{A}}, \boldsymbol{h}_m^{\mathcal{A}}))} \tag{9}$$

$$\text{where } \mathcal{X}^{\mathcal{A}} = \{x_{i,-}^{\mathcal{A}}\} \cup \{x_m^{\mathcal{A}}\}_{m=1,m\neq i}^N, \tag{10}$$

$$\psi(\boldsymbol{h}, \boldsymbol{h}') = \frac{\boldsymbol{h}^\top \boldsymbol{h}'}{\|\boldsymbol{h}\| \cdot \|\boldsymbol{h}'\|}. \tag{11}$$

Here, the MNR loss function $\ell^{MMR}(a, b, c)$ takes $\{a, b\}$ as the positive sample pair and $\{a, c\}$ as the negative sample pair(s); $\mathcal{X}^{\mathcal{A}}$ is the set of negative samples built by replacing attributes and other trials in the same batch; $\boldsymbol{h}_i^{\mathcal{A}}$ and $\boldsymbol{h}_i^{\mathcal{A},\prime}$ are all embeddings of the trial $i$ but obtained from two separate inference with different dropout masks [6]; $\boldsymbol{h}_m^{\mathcal{A}}$ are attribute embeddings of negative samples.

[5]https://www.nlm.nih.gov/research/umls/index.html
[6]We turn on the dropout hence the output embeddings will be different from two forward inferences.

*3.3.2 Task II: Context Matching.* Task I makes the attribute embeddings by Trial2Vec capable of retrieving similar trials through cosine similarity, which is useful when only partial attributes are given. For instance, in the early feasibility analysis phase, researchers can type short descriptions like the title and several keywords to search wanted trials. By contrast, for complete trials (both most attributes and contexts are known), we make use of the context-aware embeddings by Eq. (5) for computing trial similarity. Hence We design Task II, which extends Task I to the context-aware embedding $\boldsymbol{h}_i$ on the complete document $x_i$:

$$x_i = x_i^{\mathcal{A}} \cup \{x_i^{ctx}\}. \tag{12}$$

Specifically, we also replace arbitrary attribute in $x_i^{\mathcal{A}}$ to build $x_{i,-}$, then utilize MNR loss as

$$\ell_i^{CM} = \ell^{MNR}(x_i, x_i, \mathcal{X}), \tag{13}$$

$$\text{where } \mathcal{X} = \{x_{i,-}\} \cup \{x_m\}_{m=1,m\neq i}^{N}. \tag{14}$$

Unlike Eq. (15), here we take the cosine similarity on the context-aware embeddings, e.g., $\psi(\boldsymbol{h}_i, \boldsymbol{h}_{i,-})$. Task II hence enables complete trial document retrieval with detailed descriptions.

*3.3.3 Task III: Semantic Matching.* We consider more granular semantic meaning of embeddings in the clinical language. Unlike general documents, two clinical trials can be distinct when the key entities are changed. For example, two clinical trials titled "*Clinical Investigation of Safety and Performance of the Sentio System*" and "*Clinical Investigation of Safety and Performance of the OCT System*" only differ in one word. But these two trials are studying different things. In this case, it is imperative to improve the embeddings' sensitiveness towards the slight change of the key entities.

In Task III, we propose to replace key entities in attributes to build negative samples $x_i^{\mathcal{A}}$. To locate the key entities, we leverage SciSpacy [28] for entity linking, i.e., entities in raw texts are extracted and linked to UMLS to obtain their canonical names. A positive sample $x_{i,+}$ can be built by replacing the raw entities $x_{i,j}^{a}$ with their canonical names $x_{i,j}^{a,+}$, i.e., $x_{i,+}^{\mathcal{A}} = \{x_i^a\}_a \setminus \{x_i^a\} \cup \{x_{i,+}^a\}$. Also, we build negative samples by replacing $x_{i,j}^{a}$ with entities $x_{i,j}^{a,-}$ distinct from the raw illustrated by UMLS. The semantic matching loss is hence

$$\ell_i^{SM} = \ell^{MNR}\left(x_i^{\mathcal{A}}, x_{i,+}^{\mathcal{A}}, \mathcal{X}^{\mathcal{A}}\right) = -\log \frac{\exp(\psi(\boldsymbol{h}_i^{\mathcal{A}}, \boldsymbol{h}_{i,+}^{\mathcal{A}}))}{\sum_{x_m^{\mathcal{A}} \in \mathcal{X}^{\mathcal{A}}} \exp(\psi(\boldsymbol{h}_i^{\mathcal{A}}, \boldsymbol{h}_m^{\mathcal{A}}))}, \tag{15}$$

$$\text{where } \mathcal{X}^{\mathcal{A}} = \{x_{i,-}^{\mathcal{A}}\} \cup \{x_m^{\mathcal{A}}\}_{m=1,m\neq i}^{N}. \tag{16}$$

The final objective is a weighted summation of the above-mentioned tasks objectives:

$$\ell_i = \ell_i^{AM} + \beta \ell_i^{CM} + \gamma \ell_i^{SM}, \tag{17}$$

$\beta$ and $\gamma$ are hyperparameters. In our experiments, we identify setting $\beta = \gamma = 1$ gets satisfying results.

# 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 retrieval performance?
- **Exp 4.** How 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 4: 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

*4.1.1 Trial Retrieval.*
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 [36] 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

$$prec@k = \frac{\#\text{ of relevant trials in the top k results}}{k}, \tag{18}$$

$$rec@k = \frac{\#\text{ of relevant trials in the top k results}}{\#\text{ of relevant trials in all candidate trials}}. \tag{19}$$

*4.1.2 Trial Outcome Prediction.*
Although Trial2Vec is designed for dense retrieval, the learned embeddings are robust for downstream tasks. To verify this claim, we take the encoder of Trial2Vec as the backbone for trial outcome prediction. To predict the binary outcome, 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 4. 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

**Table 5: 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) |
| BioBERT | 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)** |

test set; the remaining 10% is used as the validation set for tuning and 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 [35, 36]. 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 [27]. 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.
- BioBERT [23]. It is a BERT model [12] pre-trained on large-scale biomedical corpora. We take an average pooling over all token embeddings at the last layer of BioBERT for similarity computation.
- BERT-Whitening [18, 37]. This is an unsupervised post-processing method that uses anisotropic BERT embeddings [13, 24] to improve semantic search. We take the average of last and first layer of its BERT embeddings following Su et al. [37].
- BERT-SimCSE [15]. It is a contrastive sentence representation learning method stemming from MNR loss. It simply takes other samples in batch as negative samples.

We keep all methods' embedding dimensions at 768 for a fair comparison. We start from a BERT-base model to continue pre-training on clinical domain corpora, yielding our `Trial2Vec`. We take 5 epochs with batch size 100 and the learning rate 5e-5. In the second SSL training phase, we keep the training pre-trained `Trial2Vec` by AdamW optimizer with a learning rate of 2e-5, batch size of 50, and weight decay of 1e-4. Experiments were done with 6 RTX 2080 Ti GPUs.

## 4.3 Exp 1. Complete Trial Retrieval

We conduct retrieval based on the labeled complete trial document pairs in §4.1.1. We test the performance of prec/rec@k with $k = 1, 2, 5$. Since labels are unavailable in practice, we only chose unsupervised/self-supervised baselines. Results are shown by Table 5. `Trial2Vec` outperforms all baselines with a great margin. It has around 15% improvement on each metrics than the best

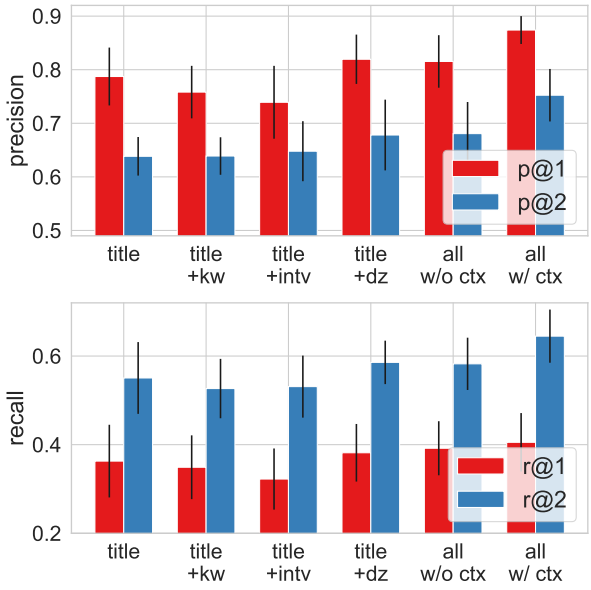


**Figure 3: Performance of `Trial2Vec` on the partial retrieval scenarios. We use a different part of the trial as queries to retrieve similar trials, including keyword *kw*, intervention *intv*, disease *dz*, context *ctx*. Error bars indicate the 95% confidence interval of results.**

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. (18). 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* [9] used in general document retrieval but invalidated in clinical trial retrieval.

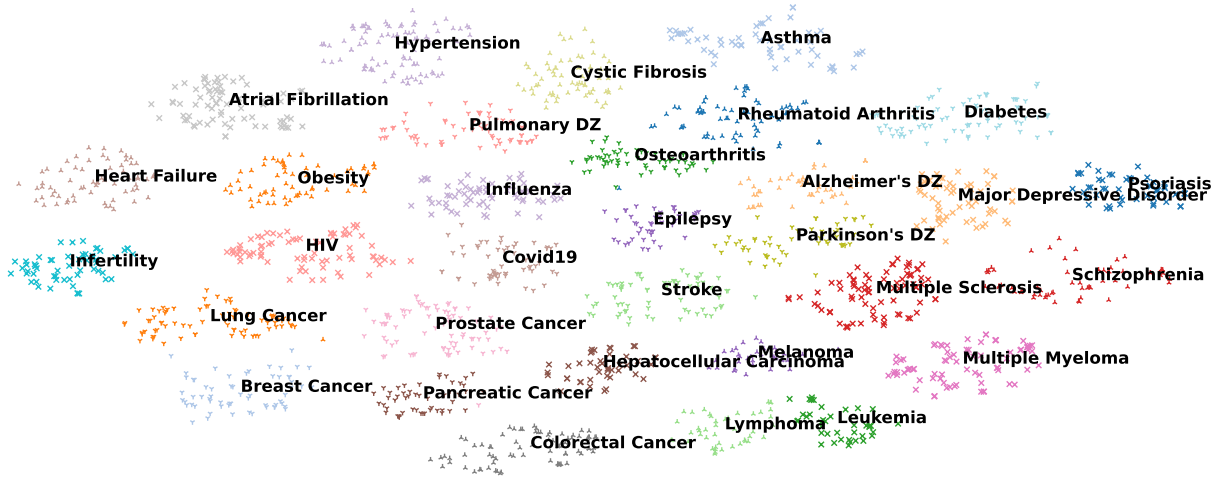


**Figure 4: 2D visualization of the trial-level embeddings obtained by Trial2Vec (dimension reduced by t-SNE). It can be seen trials are automatically classified into clusters by topic (diseases) in the embedding space. For example, a series of tumor-related trials (e.g., Breast Cancer, Pancreatic Cancer, Lymphoma, etc.) are on the bottom of the embedding space.**

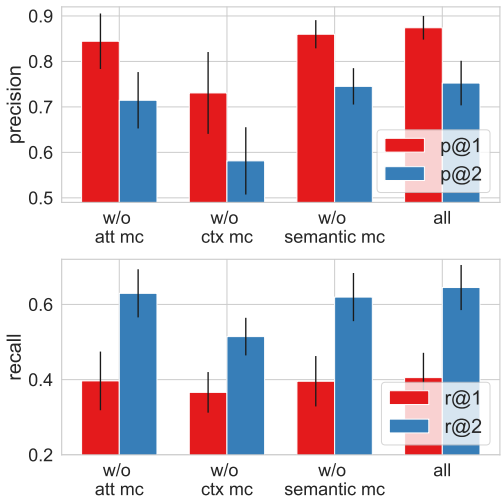


**Figure 5: Ablation study on the contribution of each Task to the final result. *att, mc, ctx* are short for attribute, matching, context, respectively. *all* indicate the full Trial2Vec that all tasks are used.**

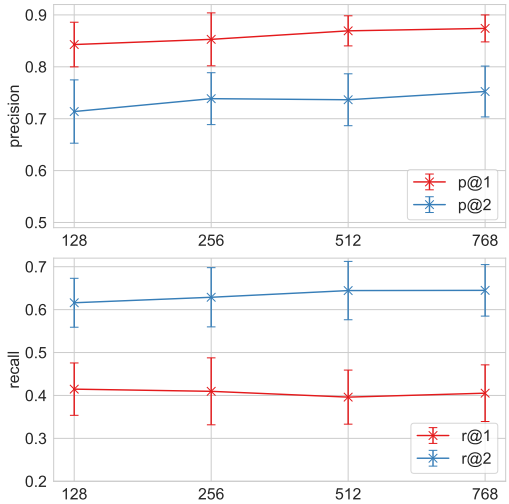


**Figure 6: Analysis of the influence of embedding dimensions on retrieval quality by Trial2Vec: embedding dim in 128, 256, 512, 768. Error bars show the 95% confidence interval.**

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 [3]. We also observed the simple negative sampling strategy of SimCSE is insufficient to learn effective long document embeddings. In comparison, Trial2Vec leverages the structure and characteristic of clinical trials to focus on the most informative attributes, with additional context-based refinement, producing embeddings superior in semantic representation.

## 4.4 Exp 2. Partial Trial Retrieval

We further investigate the partial trial retrieval scenario where users intend to find similar trials with short and incomplete descriptions. This function is enabled by Task I attribute embeddings $\boldsymbol{h}^{\mathcal{A}}$.

Results are illustrated by Fig 3. We can regard the title as the highly compressed trial descriptions like the query that users may

**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. One can browse the details of these trials by searching the NCT ID on ClinicalTrials.gov.**

| Query Trial | TF-IDF | BioBERT | 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 |

**Table 7: 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) |
| BioBERT | 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)** |

provide. We start by measuring how well Trial2Vec only utilizes the title for trial retrieval. We can see that using title is sufficient to yield comparable performance as the best baseline for complete retrieval shown in Table 5. Nonetheless, we identify that concatenating keywords or intervention with the title reduces performance. On the other hand, combining title and disease yields similar performance as involving all attributes. The additional attentive refinement performed by context embeddings further improves the performance significantly. This phenomenon signifies that the disease plays a vital role in trial similarity and is always recommended to be involved in partial trial retrieval. However, keywords/interventions might be too general sometimes such that it is encouraged to retrieve several trials with irrelevant targeted diseases. For instance, a trial named *Treatment of Latent Autoimmune Diabetes of the Adult* provides keywords including adults, Beta cell rest, Insulin secretion, etc., which are general and may deteriorate the embeddings for retrieval.

## 4.5 Exp 3. Ablation Studies on SSL Tasks and Embedding Dimension

We conducted ablation studies to measure how SSL tasks and embedding dimensions contribute to final results. Results are shown by Fig. 5, where we remove one Task for each setting and reevaluate. We can observe that Task II (context matching) is very important in trial similarity search. Without Task II, 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 without Task II 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. 6 illustrates the retrieval performance of Trial2Vec on different embedding dimensions for clinical trials. 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. 4 plots the 2D visualization of the embedding space of Trial2Vec using t-SNE [38]. There are around 2k trial embeddings $\boldsymbol{h}$ encoded by Trial2Vec (uniformly sampled from 30k 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.

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.

## 4.7 Exp 5. Trial Outcome Prediction

As described by §4.1.2, we select 244,716 trials for the experiments of trial outcome predictions. We choose the trial embeddings obtained by TF-IDF and Word2Vec with a linear layer to make binary classification. For BioBERT and Trial2Vec, we fine-tune the whole

model. Results are illustrated by Table 7. Compared with the shallow models, BERT-based methods gain better performance, which credits the deep architecture of transformers with stronger learning capability. With the two-stage pre-training and SSL training, Trial2Vec fits better to the clinical trials thus yielding superior performance on trial outcome prediction after fine-tuning than others. It sheds light ton leveraging Trial2Vec for a wide range of downstream tasks using dense trial embeddings.

### 4.8 Exp 6. Case Study

We also conducted a qualitative analysis of retrieval results by Trial2Vec and other two baselines: TF-IDF and BioBERT, by discussing two cases. Results are shown in Table 6. 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. BioBERT 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.

In summary, these two case studies show that TF-IDF and BioBERT 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. However, compared to Trial2Vec, baseline models lack important information such as the purpose of the clinical trial. The top-1 relevant clinical trial retrieved by Trial2Vec, on the other hand, provides a more similar trial than others thanks to the SSL design regarding the structure of clinical trials.

## 5 CONCLUSION

This paper investigated how to extend BERT for dense clinical trial retrieval and proposed Trial2Vec. Our method can learn from clinical trials through self-supervision, thus free of expensive labeling of clinical trial relevance. Experiments demonstrate that Trial2Vec not only gains superior performance for trial retrieval but also benefits downstream trial-related tasks like trial outcome prediction after fine-tuned. The qualitative analysis, including embedding space visualization and case studies, further verifies that Trial2Vec gets a medically meaningful understanding of clinical trials. It is promising to extend Trial2Vec for more trial-related machine learning tasks, e.g., trial summarization, in the future.

## ACKNOWLEDGEMENT

## REFERENCES


[1] 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*. 632–642.

[2] 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*.

[3] 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* (2017).

[4] 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*.

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

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big Self-Supervised Models are Strong Semi-Supervised Learners. *Advances in Neural Information Processing Systems* 33 (2020), 22243–22255.

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

[8] 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* (2021).

[9] 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* (2020).

[10] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters* 80 (2016), 150–156.

[11] 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*. 65–74.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[13] 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*. 55–65.

[14] 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.

[15] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv preprint arXiv:2104.08821* (2021).

[16] 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*. 55–64.

[17] 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* (2020).

[18] 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* (2021).

[19] 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*. 6769–6781.

[20] Tom Kenter, Alexey Borisov, and Maarten De Rijke. 2016. Siamese CBOW: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640* (2016).

[21] Tom Kenter and Maarten De Rijke. 2015. Short text similarity with word embeddings. In *ACM International on Conference on Information and Knowledge Management*. 1411–1420.

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

[23] 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 (2020), 1234–1240.

[24] 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*.

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

*arXiv:2010.11386* (2020).

[26] 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 (2021), 329–345.

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

[28] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, 319–327. https://doi.org/10.18653/v1/W19-5034 arXiv:arXiv:1902.07669

[29] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[30] 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 (2018), 111–182.

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

[32] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 539–548.

[33] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*. 3982–3992.

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

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

[36] Gerard Salton, Edward A Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Commun. ACM* 26, 11 (1983), 1022–1036.

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

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

[39] 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*. 165–174.

[40] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 353–355.

[41] 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*. 3349–3358.

[42] 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* (2020).

[43] 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*.

[44] 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* (2021).

[45] 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*. 1253–1256.

[46] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-End Open-Domain Question Answering with BERTserini. In *Conference of the North American Chapter of the Association for Computational Linguistics*. 72–77.

[47] 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*. 1154–1156.

[48] Hamed Zamani and W Bruce Croft. 2017. Relevance-based word embedding. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 505–514.

[49] 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*. 497–506.

[50] 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*. 1601–1610.