Combining Attention-based Models with the MeSH Ontology for Semantic Textual Similarity in Clinical Notes

Noushin Salek Faramarzi[*] Akanksha Dara[*] Ritwik Banerjee[*†]
[*]Department of Computer Science
[†]Institute for AI-Driven Discovery and Innovation
Stony Brook University, New York 11794-2424, USA
{nsalekfarama,adara,rbanerjee}@cs.stonybrook.edu

***Abstract*—In this study, we present several transformer-based models as well as traditional machine learning methods to detect semantic textual similarity (STS) in clinical notes. We investigate transformer models pretrained on general English as well as clinical notes, and use generic English STS datasets as a supplemental corpus to clinical notes data. Our work is based on the 2019 National NLP Clinical Challenge (n2c2). We identify and annotate six types of sentences in the clinical notes corpus, and report an ensemble method that combines attention-based contextualized embeddings with a similarity score based on the MeSH ontology obtained by computing least common ancestors of clinical terms. Our approach does not need additional clinical data for model training, while still achieving comparable Pearson's correlation coefficient of** 0.901**.**

***Index Terms*—Electronic Health Records, Natural Language Processing, Clinical Semantic Textual Similarity, Transformers, MeSH**

I. Introduction

Hospitals collect vast amounts of textual data every day that contain information critical for medical decision-making, analysis, and a variety of other healthcare applications. Clinical care is often documented in free-text narrative, which includes several types of patient information such as family history, recent medical history, and medical imaging outcome interpretations as examples that are not easily captured in the coded form [1]. Electronic health record (EHR) systems have improved healthcare efficiency, but they have also resulted in poorly organized or incorrect documentation, as well as low-quality data samples largely due to a significant amount of redundant text, errors, and incompleteness due to the frequent and growing use of templates and copy-paste in Electronic Health Record (EHR) systems [2]–[4].

Thus, it is imperative to develop solutions that can condense information while retaining its value. In this context, measuring the degree of semantic resemblance between clinical text snippets, *i.e.*, clinical semantic textual similarity (STS), can play a crucial role in alleviating redundancy and highlighting new information [5]. Further, improvements in measuring textual similarity can help in the development of other clinical applications such as clinical question answering with evidence-based retrieval, clinical text summarization, semantic search, conversational systems, and clinical decision support [3].

STS has long been recognized as a fundamentally important task in natural language processing (NLP). Consequently, several semantic evaluation (SemEval) STS tasks have been organized [6]–[11]. These, however, only dealt with general English texts. Clinical language, on the other hand, is highly specialized and domain-specific. General STS solutions, thus, cannot be readily applied in the clinical domain.

Studies in clinical STS, such as [12]–[15], are far fewer due to the scarcity of data to generate and benchmark annotated corpora. In this work, we present STS in the clinical domain, employing the framework of the 2019 National NLP Clinical Challenge (n2c2)/Open Health Natural Language Processing Consortium (OHNLP) track on clinical semantic textual similarity (ClinicalSTS) [16]. This task provides a critically needed resource, without which clinical STS would not be able to leverage the advances of modern NLP research.

Neural models for STS have often used encoders to obtain embeddings, which are abstract representations of text in a semantic vector space, followed by a regression layer to arrive at the final similarity score [17]. Such models can be pretrained by learning generalizable language representations, which are helpful for downstream NLP tasks and minimize the need of training new models for most specific tasks [18], [19]. These pretrained representations allow users to extract semantic information from enormous amounts of unlabeled text data on a range of general natural language tasks, including STS [19]. This approach, however, is not directly applicable in low-resource settings. Typically, in such cases, the approach has been to fine-tune a large pre-trained model on task-specific data. Our task, however, uses the MedSTS dataset [20] with 1,642 annotated samples that are densely packed with clinical terms. Thus, only very limited fine-tuning can be done using this corpus alone. The contextualized embeddings, on the other hand, are unlikely to succeed in the target task if the target domain differs significantly from the pretraining corpus [21]. Moreover, fine-tuning studies are potentially unstable since they rely on the pretrained encoder parameters to be relatively close to an optimal configuration for the target task [22].

Motivated by these observations, we provide empirical comparisons of several transformer models pretrained on general English corpora with those pretrained using clinical corpora. We also investigate the effect of fine-tuning on clinical data using additional in-domain corpora. Additionally, we employ a similarity metric based on the MeSH [23] ontology, and examine the impact of combining it with transformer-based language models. Our experiments demonstrate that combining the ontology-based similarity with the traditional approach (*i.e.*, a regression layer after the transformer), is competitive with other state-of-the-art results on this n2c2 clinical STS benchmark corpus.

II. Related Work

Semantic textual similarity (STS) is a task to quantitatively assess the similarity between the meaning of two text snippets. It is commonly addressed as a regression problem, with a real-value score used to reflect the degree of semantic similarity between two pieces of text.

A. Early research in semantic textual similarity

Early research in both the general and clinical domains centered on methods including lexical semantics, basic syntactic similarity, surface form matching, and alignment-based approaches [24]–[26]. The motivations behind these approaches are identification, alignment, and scoring of semantically-related words and phrases before a piece-wise aggregation of the scores. In this sense, this body of work bears resemblance with early work in machine translation.

However, due to the lack of a consistent method for merging semantic information, sentence representations were used [27], [28]. While this was an improvement over preceding work, it did not consider the context when creating distributed representations, thereby leaving considerable room for improvement. Consequently, STS and related areas of natural language processing saw multiple attempts at generating richer representation to encode the linguistic characteristics of a phrase. For STS in particular, this includes paragraph vectors [29]–[31], representation weighting and component removal [32], and convolutional deep structures [33], [34].

B. Language models and task-specific transfer learning

On the other hand, recent advances in learning sentence representation have used pretrained language models [19], [35], [36]. The *bidirectional encoder representations from transformers*, or BERT [19], employs the transformer architecture proposed by Vaswani et al. [37] to create rich sentence representations that achieve state-of-the-art results in a variety of downstream NLP tasks. Their success has lead to the development of domain-specific variants like BioBERT [38], which have found success in medical language tasks [39]. Whether for general representations or domain-specific ones, the usual approach in prior work has been to employ language models obtained through pretraining on a large amount of data, and then fine-tune the model parameters for the target task – in the spirit of transfer learning – often with the addition of a task-specific output layer.

C. Semantic similarity of clinical texts

Only recently has clinical STS has received a lot of attention, primarily through the n2c2 tasks [16], [20]. In these tasks, performance has been measured by computing the Pearson correlation coefficient (r) between text-pairs.

The clinical STS 2018 n2c2 submissions combined traditional machine learning algorithms like random forests with more recent neural architectures, including recurrent and convolutional neural networks. Pertaining to our work, of special interest is the approach taken by Chen et al. [40], who incorporate several linguistic features with deep learning models to achieve the best result ($r = 0.833$).

Subsequently, a much larger competition was held in 2019, where the top-performing approaches used state-of-the-art neural models together with pretraining and fine-tuning. The best result ($r = 0.901$) was achieved by Mahajan et al. [41], who used ClinicalBERT for multi-task learning. Their approach required a significant amount of additional training on intermediate labeled tasks, and employed several additional publicly available corpora: the SemEval 2017 Semantic Textual Similarity Benchmark (STS-B) [11], medical question entailment [42], clinical natural language entailment [39], and Quora question pairs [43]. This work also created and used two additional datasets – one on sentence-level topics, and another on drug named entity recognition. In natural language semantics, similarity and entailment are closely interlinked, so much so that general STS has been investigated purely as an entailment problem (*e.g.*, Castillo and Estrella [44]). It is therefore not surprising that training a neural model on multiple entailment datasets leads to excellent performance. Progressing along this line of work, however, makes clinical STS reliant on parallel progress along other NLP tasks.

Another competitive approach – interesting in its deviation from the traditional use of a regression layer – saw the incorporation of an ensemble algorithm into BERT, where the regression head was duplicated before applying an adapted training strategy to facilitate the focus of these multiple heads on different input patterns in the text [45]. A graph-based similarity measure was used to compute the final similarity score between two pieces of texts, to achieve $r = 0.897$.

BERT and its variants were also combined with various forms of domain knowledge embeddings. Chang et al. [46] used BERT to encode text-pairs (*i.e.*, to obtain a vector representation for each datum in the task corpus), while also constructed knowledge graphs based on the sentences to model the various concepts present in them. Finally, they used graph convolutional networks to encode these knowledge graphs. An ensemble of different language model variants for knowledge distillation, and taking a final ensemble of the language models after incorporating data augmentation and the knowledge graph encoding, yielded the best performance of $r = 0.882$. On the other hand, Xiong et al. [47] added character- and entity-level embeddings to augment BERT. Compared to the construction of knowledge graphs for each text-pair, this approach is lightweight, where the only use of

TABLE I
SIMILARITY SCORE DISTRIBUTION FOR SENTENCE-PAIRS

Dataset	**# Pairs**	[0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]
ClinSTS (train)	1,642	312	154	394	509	273
ClinSTS (test)	412	238	46	32	62	34
STS-G	28,518	3,318	2,915	4,750	9,326	8,209

external domain knowledge comes in the form of encoding medical entities by their MeSH representations.

Our approach bears resemblance with the above body of work in that we, too, employ general and domain-specific language models, and make use of additional domain knowledge. We do not, however, rely on corpora of other tasks distinct from STS (such as entailment or question-answering). Our approach also requires significantly less training because we make direct use of the MeSH ontology. We accomplish this by computing the least common ancestor (LCA) of medical entities in MeSH, instead of training neural networks to encode the entities for subsequent use in the regression layer of a neural architecture. Furthermore, we identify six distinct types of texts in the clinical STS task, and exploit this insight in our ensemble method. Our approach employs fewer additional datasets and requires less training, but nevertheless achieves state-of-the-art performance with $r = 0.901$.

III. DATA AND EVALUATION

In this paper, we use the corpus distributed for the 2019 n2c2 clinical STS task. This is a gold-standard annotated subset of the MedSTS corpus, which is a large collection obtained from de-identified clinical notes of patients receiving their primary care at Mayo Clinic [3]. The annotated subset for the 2019 n2c2 clinical STS task consists of 1K sentence-pairs, plus the dataset from the 2018 clinical STS task. In both years, two clinical experts independently annotated each sentence-pair to provide a score in the continuous range 0 (complete dissimilarity) to 5 (complete semantic equivalence). The inter-rater agreement for both rounds is given by the weighted Cohen's kappa scores, $\kappa = 0.6$ and $\kappa = 0.67$, respectively. The final dataset comprises 1,642 sentence-pairs in the training set, and an additional 412 sentence-pairs in the test set.

Given the small size of the training data, we also use non-clinical STS data from the SemEval shared tasks 2012-2017 [6]–[11]. These dataset provide pairs where understanding semantic similarity requires the identification of multi-word expressions, recognition of named entities, or accessing encyclopedic knowledge. But they offer relatively poor coverage of other semantic challenges such as resolving ambiguous synonymy based on context, active/passive voice, the scope of operators, and other lexical variations. This weakness of existing STS corpora is noted by Marco et al. [48], who develop a collection of about 10K pairs of *sentences involving compositional knowledge* (SICK) to fill this gap. In one leg of our experiments, we combine the SemEval STS corpora with the SICK collection to investigate the performance of transformer-based language models (see Sec. IV-B). This yields a much larger collection of 28,518 sentence-pairs with gold-standard annotations, which we dub STS-G. Table I shows the distribution of the gold-standard semantic similarity scores in the datasets we use.

For evaluation, we adopt the same measure as used in the original n2c2 clinical STS task: Pearson correlation coefficient (r) between the predicted similarity scores and the average of the two clinical expert judgments.

In addition to the sentence-pair data, we also utilize the MeSH ontology (for details, see Section IV-C). This is similar in spirit to the use of external dictionaries in – among other healthcare applications of NLP – clinical decision support systems. However, the use of medical dictionaries require careful engineering of heterogeneous data sources, increasing the possibility of errors in a pipeline. It also necessitates pre- and post-processing steps to handle complex expressions like multi-word entities and non-standard acronyms. Ontology-based approaches, while comparable to the use of dictionaries in some ways, focus on the use conceptual knowledge representation, often categorized by fine-grained semantic types. This enables the exploitation of both hierarchical and non-hierarchical relationships between entities. An ontology like MeSH, for instance, instantly allows for processing information of the type "Aspirin *is-a* antipyretic" – a capability typically not found in dictionaries. Overall, when it comes to incorporating domain knowledge beyond a given corpus, a structured ontology offers both theoretical and practical advantages over the use of dictionaries. This motivates our use of MeSH over other available resources.

IV. EXPERIMENTAL APPROACHES AND RESULTS

To understand clinical STS, our experimental approach is three pronged: (A) develop an ensemble model that combines neural language models with several text similarity metrics, (B) investigate combinations of transformer-based language models in conjunction with regression, and (C) use the MeSH ontology to directly inject similarity scores for entity-pairs into the computation of semantic similarity for pairs of texts. This third approach allows for the extrapolation of similarities across entities from the training set.

A. Combining Deep and Shallow Learning Techniques

To obtain the semantic similarity score between sentence pairs, we first look into combining traditional machine learning algorithms with state-of-the-art neural models. We use three language models based on the transformer architecture. First, we employ a universal language model that combined the pre-trained bidirectional transformer language model BERT with multi-task learning. This model, called multi-task deep neural network (MT-DNN) [49], has been shown to outperform BERT in several natural language understanding benchmark tasks. Further, we use SciBERT [50] and BioBERT [38], which are pretrained on the same masked language modeling and next sentence prediction tasks as the original language model developed by Devlin et al. [19], but on vast amounts of biomedical text.

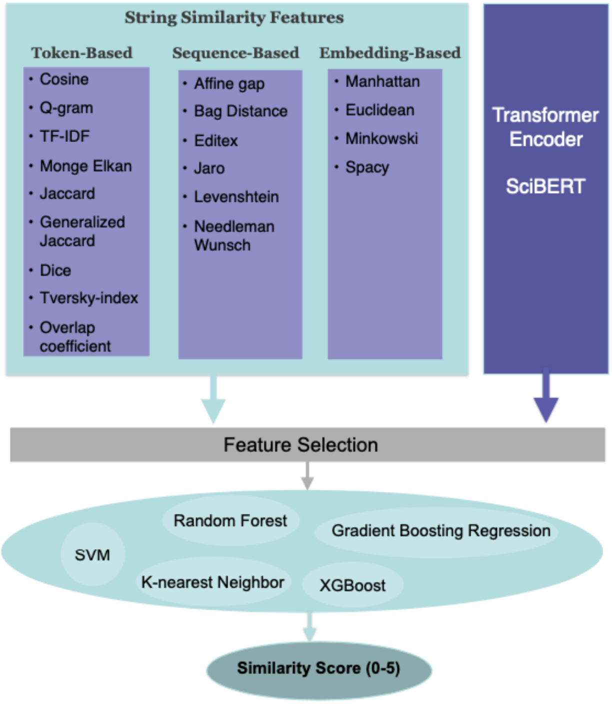

Fig. 1. Combination of Deep and Shallow Learning Techniques.

One common hurdle in natural language tasks is that a single entity may be mentioned in multiple synonymous lexical forms. Entity linking is the process of connecting all the textual mentions of an entity to a canonical representation [51]. The typical approach, especially when knowledge bases of such canonical representations exist, has been to use thesauri to perform entity linking. We adopt this approach, and use the *unified medical language system* (UMLS) metathesaurus [52][1]. To link medical entities, we replace all medical terms with their UMLS preferred terms[2] We implement the mapping of each entity to its preferred term, we use the MetaMap tool [53].

As depicted in Fig. 1, our input is a collection of several string similarity measures together with an encoding for the sentence pair. For the string similarity measures, we compute multiple token-based, sequence-based, and vector-space metrics. Next, we use SciBERT to obtain the vector representation of each sentence in a pair, and concatenate them using the separator [SEP] token. We thus obtain a single vector representation for each sentence pair. This representation, together with the similarity measures, form the input vector

We perform feature selection on this input vector, by training a Lasso model and discarding the weakly correlated features with vanishing coefficients, since these features make no contribution to the prediction. On the feature vector thus obtained, we test a variety of models: support vector regression, random forest, gradient boosting regression tree, k-nearest neighbor, and extreme gradient boosting.

TABLE II
TRANSFORMER-BASED MODELS AND THEIR TRAINING CORPORA

Language model	Pretraining data
RoBERTa	Wikipedia + Books Corpus
BioBERT	Wikipedia + Books Corpus + PubMed + PMC
ALBERT	English Wikipedia + Books Corpus
PubMedBERT	PubMed abstracts and articles
SciBERT	Semantic Scholar
XLNet	Books Corpus + Wikipedia + ClueWeb 2012-B
Bio_ClinicalBERT	MIMIC III database

The ensemble of these traditional models achieve a better score than any individual traditional model. Additionally, the ensemble of these traditional models with the neural language models perform better than the neural language models alone. This leg of our experiments comprise many different experiments with various model parameters (*e.g.*, the regularization parameter for feature selection using Lasso regression) and choice of language model. For the sake of brevity, we omit the details and present the best model, where we use SciBERT and achieve $r = 0.868$.

B. Learning similarity through Transformer-based models

Our first approach, as described above in Section IV-A, suffers largely due to the small size of the training data. To alleviate this concern, we employ the STS-G dataset (see Section III) and investigate several transformer-based language models, both domain-specific and general.

Due to their ability to internally deploy attention mechanisms, transformer-based models are capable of simulating syntactic and semantic constructs in language. It is perhaps due to this reason that such models find success in a variety of downstream tasks with limited fine-tuning, as long as the task employs language in a similar domain and/or genre. For different domains, however, the pretraining data is known to have significant effects on the model, since the vocabulary tends to differ as well. Consequently, several models have been developed on the same transformer architecture, but with different data for pretraining. Next, we provide a brief discussion of these models. We would also like to draw attention to Table II, which displays the corpora used for pretraining these models[3].

Language Models: RoBERTa, ALBERT, and XLNet are models that were trained on large amounts of general English language data. RoBERTa [55] uses the same transformer architecture, but builds on BERT with significant changes in hyperparameter choices. It also removes the next sentence prediction from the pretraining process, and trains with much larger batch sizes and learning rates. ALBERT [56] is a lightweight BERT model, in the sense that it uses parameter-reduction techniques for faster training. It is also known to better capture inter-sentence semantics than the original BERT

[1] UMLS brings together several biomedical vocabularies to enable interoperability. For instance, the entity “oral anticoagulant” and all its commonly used variations are mapped to a single canonical concept in UMLS.

[2] Every UMLS concept has a unique concept identifier, as well as a canonical *preferred term*.

[3] For details about the Books Corpus, see Zhu et al. [54].

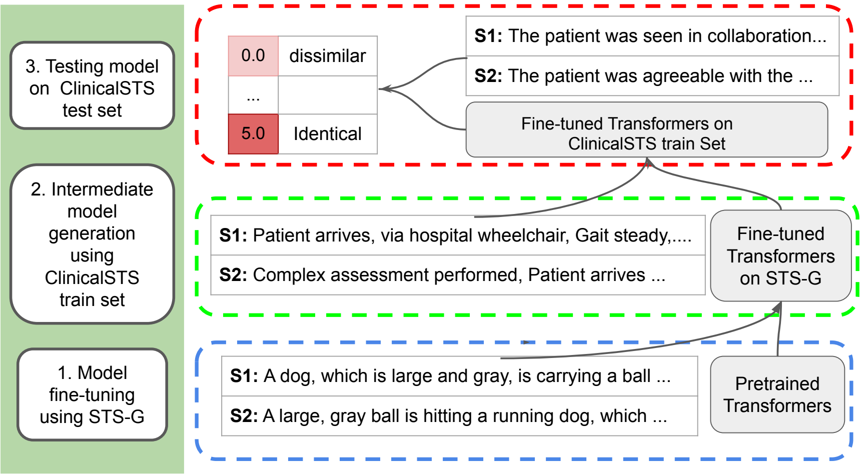
3. Testing model on ClinicalSTS test set
2. Intermediate model generation using ClinicalSTS train set
1. Model fine-tuning using STS-G
0.0
dissimilar
...
5.0
Identical
S1: The patient was seen in collaboration...
S2: The patient was agreeable with the ...
Fine-tuned Transformers on ClinicalSTS train Set
S1: Patient arrives, via hospital wheelchair, Gait steady,....
S2: Complex assessment performed, Patient arrives ...
Fine-tuned Transformers on STS-G
S1: A dog, which is large and gray, is carrying a ball ...
S2: A large, gray ball is hitting a running dog, which ...
Pretrained Transformers

Fig. 2. The two-stage training strategy for clinical STS.

model. XLNet [57] is a transformer-based model that uses autoregressive methods to learn bidirectional contexts in the language. Unlike the other models described here, XLNet uses a data permutation method for training instead of the more standard masked language modeling. It thereby avoids data corruption and reconstruction, and better captures the dependency between word positions.

With the exception of SciBERT, which is trained on research articles from multiple scientific disciplines, the other models we use are specific to the medical domain. SciBERT [50] is trained using an architecture identical to BERT's base model. It, however, uses different initialization weights because it uses its own custom vocabulary. This vocabulary includes a large number of tokens that are not relevant to the language used in clinical notes. Moreover, its restriction to 30K subwords also plays a role in preventing several medical terms from being modeled. BioBERT [38] uses the same architecture as BERT's base model, and is trained with identical weights. It only differs in the additional domain-specific training with medical research publications. Bio_ClinicalBERT [58] is trained on notes from the MIMIC III corpus [59], a large collection containing EHRs. In terms of domain-specific pretraining, it is a likely candidate to deliver excellent performance on a task that uses clinical notes. The size of the pretraining corpus, however, is smaller than the ones used by most other language models. Another domain-specific model is PubMedBert [60], which constructs a domain-specific vocabulary with 3.1 billion words (21GB of data, compared to the 16 GB used by BERT). The abstracts and full biomedical articles are used to train a BERT (base) architecture from scratch.

Before moving toward describing our approach to training and fine-tuning various models, we would like to highlight that most transformer-based models provide two encoder architectures: “base” and “large”. The main difference between them is the number of layers. For example, the BERT-base model features 12 layers of transformer encoder layers, 768 hidden units in each layer, and 12 attention heads, while the BERT-large consists of 24 transformer layers with a hidden size of 1,024 and 16 attention heads.

C. *Learning similarity from the MeSH ontology*

Training Strategy: We use the transformer-based models to learn distributed sentence-level representations from sentence pairs. The linear regression layer then uses these distributed encodings to obtain a similarity score between 0 and 5. Our training comprises two steps. First, we take a pretrained model and further train it on the STS-G corpus, which is a dataset for general semantic similarity. Second, we fine-tune the model using the training data from the clinical STS dataset. In both steps, hyperparameters are optimized by 5-fold cross validation. We use a learning rate $\eta = 10^{-5}$. Our experiments varied the batch size between 3 and 4, and the number of epochs between 2, 4, and 8.

To create distributed representations, each transformer is deployed. To train our clinical STS models, we use a two-step technique as shown in Fig. 2. In the first step, the STS-G corpus was employed to fine-tune an intermediate STS model. Using the ClinicalSTS corpus, the intermediate model was fine-tuned even further in step 2. The fine-tuned model from the second phase was used for final testing. We employed 5-fold cross-validation to optimize hyperparameters in both step 1 and step 2 of training. We use $\eta = 10^{-5}$ as a learning rate and tested batch size selected as 3 and 4 and the number of epochs are also tested as 2,4, and 8. The epoch number, batch size, and learning rate were all adjusted based on the cross-validation findings.

We use the PyTorch-based models from the HuggingFace transformers in our experiments [61], [62]. The RoBERTa-large model outperforms all the others with a Pearson correlation coefficient of 0.896. For RoBERTa and XLNet – both pretrained on general English language data – the large models perform better, as expected. Surprisingly, BERT's large

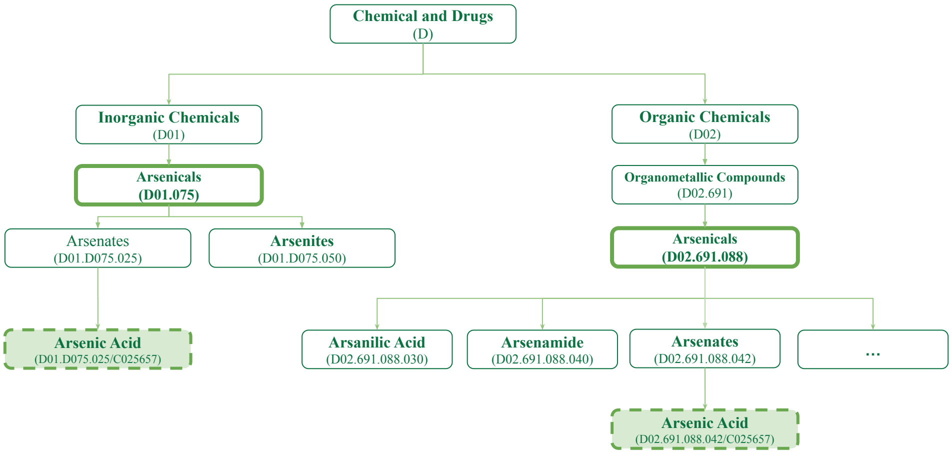

Fig. 3. The MeSH descriptors are arranged in logical hierarchical groups (families), and a term may appear in multiple nodes/locations in the hierarchy, as shown for “Arsenic Acid” here. Semantically similar terms are closer together in this structure, which motivates us to compute the least common ancestor of term-pairs as a way of computing the similarity of paired sentences in the clinical STS dataset.

and base models show nearly equivalent results (0.867 and 0.868, respectively). In every case, the models pretrained using general English corpora (both large and base) outperformed their corresponding domain-specific counterparts pretrained using clinical notes from the MIMIC-III dataset. We conjecture this is due to the relatively (and significantly) smaller size of the MIMIC-III data. These results, along with the results of our other experiments, are shown in Table IV.

The ClinicalSTS dataset indicates two features that we take into account when approaching this task. First, due to inadequate training data, constructing reliable learning models is difficult without augmentation with additional training data. Second, medical entities have a natural hierarchical grouping, based on their pharmacologic action or their chemical structure (drug entities), their biological function (anatomical or microbial entities), etc. These relative differences and similarities cannot be resolved simply by entity linking. There is an obvious need to include a domain-specific ontology to incorporate this knowledge.

To this end, we utilize *Medical Subject Headings* (MeSH) [23], a medical ontology produced by the National Library of Medicine. It is used to index and catalog medical documents based on key concepts present in those documents (*e.g.*, research articles in the MEDLINE database[4]. MeSH offers a controlled vocabulary, much like UMLS, for biomedical concepts. It also offers an extensive hierarchy of such concepts in the form of *headings* (also called “major headings” or “descriptors”). The hierarchy consists of 16 major branches (including “Anatomy”, “Organisms”, “Diseases”, “Chemicals and Drugs”, etc.).

In spite of the availability of such knowledge bases as MeSH or UMLS, the nearness of two medical terms is not always obvious. Fig. 3 shows how these headings are organized into logical hierarchical groups, and that a single term (Arsenic Acid) can be a member of multiple groups in the hierarchy, where each group membership is valid due to some specific property of the term.

Clinical sentence types: In no small part due to the above reasons, determining the extent of semantic similarity is particularly difficult for clinical texts. For clinical notes, however, we observe that the text may be viewed as belonging to a particular category, based on the type of information it conveys (*e.g.*, diagnosis, discharge). We thus tag each sentence with one of six labels. Next, we explain these types along with example sentence-pairs (S1, S2) and their semantic similarity scores ($\sigma \in [0, 5]$) from the clinical STS dataset.

1. **MED**. All medications, either prescribed or over the counter. It has information such as the medication name, dose, route, and frequency.
   (S1) *Oxycodone [ROXICODONE] 5 mg tablet 0.5-1 tablets by mouth every 4 hours as needed.*
   (S2) *Pantoprazole [PROTONIX] 40 mg tablet enteric coated 1 tablet by mouth Bid before meals.* ($\sigma = 1$)
2. **EDU**. Contains information that allows patients and their caregivers to understand their disease and treatment plan.
   (S1) *The patient's caregiver was ready to learn without barriers and understanding of the plan provided.*
   (S2) *The patient will verbalize understanding of splint wear and care following one treatment session.* ($\sigma = 1.25$)
3. **ADM/DISCH**. These include, but are not limited to, admission status (inpatient, ambulatory, etc.) for admitting diagnosis, information on the attending physician, vital sign parameters, known or observed allergies/reactions.
   (S1) *Patient appears comfortable, Patient cooperative, alert, Oriented to person, place and time.*

[4] https://www.nlm.nih.gov/medline/index.html

(S2) *History obtained from patient, Patient appears, anxious, Patient cooperative, alert, Oriented to person, place and time.* ($\sigma = 3.1$)

4. **REC**. Includes administrative and billing data, patient demographics, progress notes, allergens, allergies, radiology images, lab and test results.
   (S1) *The lesion was excised from within the subcutaneous tissue down to fascia.*
   (S2) *The larynx was examined, specifically the supraglottis, true vocal folds, and subglottis.* ($\sigma = 0$)
5. **DIAG**. Conveys information about the signs and symptoms of the patient.
   (S1) *There is no lower extremity edema present bilaterally.*
   (S2) *There is a 2+ radial pulse present in the upper extremities bilaterally.* ($\sigma = 1$)
6. **CON**. Conveys information about consent before conducting an intervention or to disclose personal information.
   (S1) *The above was discussed with the patient, and she voiced understanding of the content and plan.*
   (S2) *Patient was provided with written patient education materials and she relates a good understanding of the self-management program.* ($\sigma = 1.5$)

Ontological similarity based on "least common ancestor": To determine how closely related two terms are, we investigate a remarkably simple method of directly and explicitly utilizing the tree-structure of the MeSH ontology: the *least common ancestor* (LCA) algorithm [63]. This is in contrast with other approaches that create abstract embeddings from knowledge graphs or ontologies (*e.g.*, Chang et al. [46] and Xiong et al. [47]). Our motivation behind exploring this is obvious, since less similar terms will be further apart (and hence have a common ancestor further away from their own descriptor locations). Highly similar terms, on the other hand, will have a common ancestor closer to their own descriptors. As noted earlier, a term may appear in multiple locations within the MeSH hierarchy, so we iterate over all possible tree-pairs corresponding to the term-pairs obtained from the two sentences S1 and S2.

Algorithm 1 Collect all the tree nodes corresponding to terms appearing in MeSH

```
for each (S1, S2) in sentences do
    phraseIDs1, phraseIDs2 = set of all possible IDs of MeSH
    searchable phrases in S1, S2
    for phrase_A ∈ phraseIDs1 do
        for phrase_B ∈ phraseIDs2 do
            if $(phrase\_A \neq phrase\_B)$ then
                trees_A = set of all the tree nodes where phrase
                A is contained in the descriptor
                trees_B = set of all the tree nodes where phrase
                B is contained in the descriptor
                LCA = maxLCA(trees_A, trees_B)
            end
        end
    end
end
```

Algorithm 2 maxLCA(tree_set_A, tree_set_B)

```
Result: The maximum LCA score across all tree-pairs
max_LCA = 0
for tree_A ∈ tree_set_A do
    for tree_B ∈ tree_set_B do
        max_LCA = max(max_LCA, $\sigma_{LCA}$(tree_A, tree_B))
    end
end
return max_LCA
```

We begin by querying MeSH for all the terms in the training set of the clinical STS dataset, and then locally storing those entities (term, MeSH unique ID, tree locations). For example, the word "via" appears in multiple MeSH entities, each having its own unique ID, and these IDs can appear in several locations in the ontology tree. Clearly, not all words are relevant for the computation of semantic similarity from the clinical/medical perspective. So, we extract noun phrases from each sentence, and check whether the phrase is contained in at least one MeSH entry. If so, we collect all MeSH IDs associated with this phrase. We use ScispaCy [64] to extract noun phrases, and iterate over these phrases to obtain the corresponding tree nodes as shown in Algorithm 1.

We then compute LCA across the two sets of tree nodes, and retain the maximum over all the pairs, as shown in Algorithm 2. Each LCA score is based simply on a traversal of the two tree paths, where the score is initially set to 0, and incremented for each matched node. The pseudocode for this computation is shown in Algorithm 3.

To implement this approach, we use E-utilities [65], which is the public API to the NCBI Entrez system, giving access to all Entrez databases, including PubMed, PMC, Gene, Nuccore, and Protein. This API is a collection of eight server-side programs that accept a fixed URL syntax for searching, linking, and retrieving data. We specifically use two APIs available as part of E-utilities: the combination ESearch (send a text query to a single Entrez database) and EFetch (retrieve the full record for a MeSH ID). Upon sending a text query, we get a list of all the MeSH IDs where that text (*i.e.*, word or phrase) appears. We then use EFetch to obtain the complete

Algorithm 3 σ_{LCA}(tree_A, tree_B)

```
Result: LCA score for a pair of trees
LCA = 0   for i = 0 to min(len(tree_A), len(tree_B)) do
    if $tree\_A[i] == tree\_B[i]$ then
        ++LCA
    end
    else
        break
    end
end
return LCA
```

TABLE III
SENTENCE TYPE DETECTION: CLASSIFICATION RESULTS ON THE CLINICAL STS TEST SET

Type	Precision	Recall	F_1	Support
ADM/DISCH	1.00	1.00	1.00	24
MED	0.99	0.99	0.99	99
REC	0.90	0.91	0.90	127
DIAG	0.91	0.91	0.91	137
CON	0.97	0.98	0.98	113
EDU	1.00	0.95	0.98	42

records for each of these MeSH IDs. In particular, we use tree numbers in the query's result, since it shows us the subtrees where the phrase belongs. The relevance of this detail can be seen in Fig.3, where the chemical category "Arsenicals" appears at two separate nodes in the tree. The MeSH tree number is shown in parenthesis for each node.

Directly using the LCA computations with the MeSH ontology is an attractive option to measure clinical STS, due to the simplicity of this approach. However, our analysis reveals that this approach performs well only in the **MED** type sentences. The other five types often do not contain medication information, and therefore, most terms in such sentences understandably do not appear in the MeSH ontology.

Any use of the sentence type knowledge, of course, relies on being able to accurately classify sentences from clinical notes as belonging to one of the six types. We experiment with several multi-label classification algorithms, including random forest classifier, support vector machines, multinomial naïve bayes, and logistic regression. We use the Scikit-learn library for these experiments [66]. With an accuracy of 0.91, linear SVC outperforms all the other classifiers. The precision, recall, F_1 score, and support across all six sentence types in this sentence-type detection are shown in Table III.

Among all the transformer-based models we investigated, RoBERTa-large showed the best result ($r = 0.896$), as shown in Table IV. To combine the advantages of the MeSH ontology with the deep neural models, we thus consider only the RoBERTa-large model for further experiments. Adding the LCA score as an additional feature (and thereby increasing the dimension of the input vector for a sentence-pair by 1) does not yield any significant improvement in the final Pearson correlation coefficient. Looking at the dataset in terms of the six sentence types, however, we discover that for the MED sentences, the LCA score is a better predictor of the regression task than RoBERTa-large. Therefore, we combine the best of both worlds, and use the LCA score for **MED** sentences while still employing the best transformer-based model, RoBERTa-large, to predict the similarity of other sentence types. We repeat this ensemble for the other five sentence types as well, but observe no significant difference in the Pearson correlation coefficient values on the test set.

TABLE IV
PERFORMANCE COMPARISON ON THE CLINICAL STS TEST SET.

Model	Pearson
BERT-base-uncased	0.868
BERT-large-uncased	0.867
RoBERTa-base	0.879
RoBERTa-large	0.896
ALBERT-base-v2	0.870
XLNet-base-cased	0.867
XLNet-large-cased	0.875
Bio_ClinicalBERT	0.868
BioBERT-v1.1	0.856
PubMedBERT-base-uncased-abstract	0.885
PubMedBERT-base-uncased-abstract-fulltext	0.885
SciBERT	0.866
Our best performing model(RoBERTa-large), LCA	0.901
Multi Task Learning, ClinicalBERT [41]	0.901
M-heads, BERT [45]	0.883
CNN, BERT-based transformers [67]	0.896
GCN-based graph encoders, BERT-based transformers [46]	0.882
Character-level information, BERT [47]	0.868

Table IV shows the final prediction performance of our models, along with the best results from prior work on this task. Our combination of LCA with RoBERTa-large (applied posterior to sentence-type classification on the test set) achieves an overall score $r = 0.9011$, narrowly surpassing[5] the previous state-of-the-art ($r = 0.9010$) on this clinical STS dataset [41], and significantly outperforming many others.

V. CONCLUSION

In this study, we look at three different approaches to the Clinical STS challenge. First, we use an ensemble machine learning framework based on string similarity measures to account for a variety of explainable features, such as cosine similarity of domain-specific embeddings. Second, we design a system that can use several transformer algorithms and present transformer-based models for evaluating clinical STS. In our work, RoBERTa with the large architecture outperforms the other attention-based language models investigated. Finally, we identify a remarkably lightweight and simple approach of incorporating the MeSH ontology for this task, where the lowest common ancestor (LCA) of medical terms in the hierarchy is computed and used as a signal for semantic similarity.

Our work also offers additional insight into the nature of clinical notes, from the perspective of computational linguists. We observe that sentences in clinical notes can be broadly categorized based on the type of clinical information conveyed. In light of this, we label the sentences in the clinical STS corpus as one of six different categories. Further, we find that even if two sentences are similar in terms of syntactic structure and the non-medical vocabulary present in them, their clinical meaning and significance may be vastly different (*e.g.*, because the drugs mentioned are different). Based on these insights – and the intuition that the active ingredient in the drug would be the most important component in determining the difference between two sentences – we employ the MeSH ontology and compute the semantic similarity of medical terms using the lowest common ancestor (LCA) of the terms in the MeSH type

[5] This difference between our best result and the best result reported by Mahajan et al. [41] is, however, not statistically significant.

hierarchy. We find that the approach based on using the MeSH hierarchy works extremely well on sentences that explicitly mention pharmaceutical products and their dosage, while the embeddings obtained from training attention-based language models on domain-specific data perform better for any other clinical texts. Consequently, by combining the LCA measure and the best performing language model, we achieve state-of-the-art result, predicting the similarity scores of human clinical experts with a Pearson's correlation $r = 0.9011$.

Acknowledgments

Noushin Salek Faramarzi would like to thank the U.S. Department of Education for its support through the GAANN (Graduate Assistantship in Areas of National Need) Fellowship.

References

[1] S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead, and K. B. Johnson, "Data from clinical notes: a perspective on the tension between structure and flexible documentation," *Journal of the American Medical Informatics Association*, vol. 18, no. 2, pp. 181–186, 2011.

[2] T. Kuhn, P. Basch, M. Barr, T. Yackel, and M. I. C. of the American College of Physicians*, "Clinical documentation in the 21st century: executive summary of a policy position paper from the American College of Physicians," *Annals of internal medicine*, vol. 162, no. 4, pp. 301–303, 2015.

[3] Y. Wang, N. Afzal, S. Fu, L. Wang, F. Shen, M. Rastegar-Mojarad, and H. Liu, "MedSTS: a resource for clinical semantic textual similarity," *Language Resources and Evaluation*, vol. 54, no. 1, pp. 57–72, 2020.

[4] X. Yang, X. He, H. Zhang, Y. Ma, J. Bian, Y. Wu *et al.*, "Measurement of semantic textual similarity in clinical texts: comparison of transformer-based models," *JMIR medical informatics*, vol. 8, no. 11, p. e19735, 2020.

[5] S. Liu, Y. Wang, and H. Liu, "Selected articles from the BioCreative/OHNLP challenge 2018," pp. 1–3, 2019.

[6] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," in ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 2012, pp. 385–393.

[7] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "* SEM 2013 shared task: Semantic textual similarity," in *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, 2013, pp. 32–43.

[8] E. Agirre, C. Banea, C. Cardie, D. M. Cer, M. T. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe, "SemEval-2014 Task 10: Multilingual Semantic Textual Similarity." in *SemEval@ COLING*, 2014, pp. 81–91.

[9] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea *et al.*, "Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 2015, pp. 252–263.

[10] E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez Agirre, R. Mihalcea, G. Rigau Claramunt, and J. Wiebe, "Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation," in *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics), 2016.

[11] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," *arXiv preprint arXiv:1708.00055*, 2017.

[12] G. Soğancıoğlu, H. Öztürk, and A. Özgür, "BIOSSES: a semantic sentence similarity estimation system for the biomedical domain," *Bioinformatics*, vol. 33, no. 14, pp. i49–i58, 2017.

[13] Y. Xiong, S. Chen, H. Qin, H. Cao, Y. Shen, X. Wang, Q. Chen, J. Yan, and B. Tang, "Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 1–7, 2020.

[14] Q. Chen, J. Du, S. Kim, W. J. Wilbur, and Z. Lu, "Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–10, 2020.

[15] A. Wong, J. M. Plasek, S. P. Montecalvo, and L. Zhou, "Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 38, no. 8, pp. 822–841, 2018.

[16] Y. Wang, S. Fu, F. Shen, S. Henry, O. Uzuner, H. Liu *et al.*, "The 2019 n2c2/ohnlp track on clinical semantic textual similarity: overview," *JMIR Medical Informatics*, vol. 8, no. 11, p. e23375, 2020.

[17] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[18] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] Y. Wang, N. Afzal, S. Liu, M. Rastegar-Mojarad, L. Wang, F. Shen, S. Fu, and H. Liu, "Overview of the BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity," *Proceedings of the BioCreative/OHNLP Challenge*, vol. 2018, 2018.

[21] X. Han and J. Eisenstein, "Unsupervised domain adaptation of contextualized embeddings for sequence labeling," *arXiv preprint arXiv:1904.02817*, 2019.

[22] J. Phang, T. Févry, and S. R. Bowman, "Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks," *arXiv preprint arXiv:1811.01088*, 2018.

[23] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[24] F. Šarić, G. Glavaš, M. Karan, J. Šnajder, and B. D. Bašić, "Takelab: Systems for measuring semantic text similarity," in ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 2012, pp. 441–448.

[25] D. Bär, C. Biemann, I. Gurevych, and T. Zesch, "Ukp: Computing semantic textual similarity by combining multiple content similarity measures," in ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 2012, pp. 435–440.

[26] S. Jimenez, C. Becerra, and A. Gelbukh, "Soft cardinality: A parameterized similarity function for text comparison," in ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 2012, pp. 449–453.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[28] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[29] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[30] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[31] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," *arXiv preprint arXiv:1703.02507*, 2017.

[32] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *International conference on learning representations*, 2017.

[33] Y. Shao, "Hcti at semeval-2017 task 1: Use convolutional neural network to evaluate semantic textual similarity," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 130–133.

[34] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 2333–2338.

[35] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[36] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *arxiv preprint*, 2018.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[38] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[39] A. Romanov and C. Shivade, "Lessons from natural language inference in the clinical domain," *arXiv preprint arXiv:1808.06752*, 2018.

[40] Q. Chen, J. Du, S. Kim, W. J. Wilbur, and Z. Lu, "Combining rich features and deep learning for finding similar sentences in electronic medical records," *Proceedings of the BioCreative/OHNLP Challenge*, pp. 5–8, 2018.

[41] D. Mahajan, A. Poddar, J. J. Liang, Y.-T. Lin, J. M. Prager, P. Suryanarayanan, P. Raghavan, C.-H. Tsou *et al.*, "Identification of semantically similar sentences in clinical notes: Iterative intermediate training using multi-task learning," *JMIR medical informatics*, vol. 8, no. 11, p. e22508, 2020.

[42] A. B. Abacha and D. Demner-Fushman, "Recognizing question entailment for medical question answering," in *AMIA Annual Symposium Proceedings*, vol. 2016. American Medical Informatics Association, 2016, p. 310.

[43] L. Sharma, L. Graesser, N. Nangia, and U. Evci, "Natural language understanding with the quora question pairs dataset," *arXiv preprint arXiv:1907.01041*, 2019.

[44] J. Castillo and P. Estrella, "SAGAN: an approach to semantic textual similarity based on textual entailment," in ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 2012, pp. 667–672.

[45] K. Kades, J. Sellner, G. Koehler, P. M. Full, T. E. Lai, J. Kleesiek, and K. H. Maier-Hein, "Adapting Bidirectional Encoder Representations from Transformers (BERT) to Assess Clinical Semantic Textual Similarity: Algorithm Development and Validation Study," *JMIR medical informatics*, vol. 9, no. 2, p. e22795, 2021.

[46] D. Chang, E. Lin, C. Brandt, R. A. Taylor *et al.*, "Incorporating Domain Knowledge Into Language Models by Using Graph Convolutional Networks for Assessing Semantic Textual Similarity: Model Development and Performance Comparison," *JMIR Medical Informatics*, vol. 9, no. 11, p. e23101, 2021.

[47] Y. Xiong, S. Chen, Q. Chen, J. Yan, B. Tang *et al.*, "Using Character-Level and Entity-Level Representations to Enhance Bidirectional Encoder Representation From Transformers-Based Clinical Semantic Textual Similarity Model: ClinicalSTS Modeling Study," *JMIR Medical Informatics*, vol. 8, no. 12, p. e23357, 2020.

[48] M. Marco, B. Luisa, B. Raffaella, M. Stefano, Z. Roberto *et al.*, "SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," in *Proc. SemEval*, 2014, pp. 1–8.

[49] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," *arXiv preprint arXiv:1901.11504*, 2019.

[50] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[51] W. Shen, J. Wang, and J. Han, "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 443–460, 2015.

[52] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[53] A. R. Aronson, "Metamap: Mapping text to the umls metathesaurus," *Bethesda, MD: NLM, NIH, DHHS*, vol. 1, p. 26, 2006.

[54] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015, pp. 19–27.

[55] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[56] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[57] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[58] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[59] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[60] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[61] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[62] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[63] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, "On finding lowest common ancestors in trees," *SIAM Journal on computing*, vol. 5, no. 1, pp. 115–132, 1976.

[64] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. [Online]. Available: https://www.aclweb.org/anthology/W19-5034

[65] J. Kans, "Entrez direct: E-utilities on the UNIX command line," in *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US), 2022.

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[67] Q. Chen, A. Rankine, Y. Peng, E. Aghaarabi, Z. Lu *et al.*, "Benchmarking Effectiveness and Efficiency of Deep Learning Models for Semantic Textual Similarity in the Clinical Domain: Validation Study," *JMIR Medical Informatics*, vol. 9, no. 12, p. e27386, 2021.