Exploring Clinical NLP with Pre-trained Language Models

Qiuhao Lu
Department of Computer Science
University of Oregon
Eugene, OR, USA
luqh@cs.uoregon.edu

Abstract

Pre-trained Language Models (PLMs) have been one of the fundamental components of natural language processing techniques over the past few years, and have proven their efficacy across a wide range of applications. In the clinical field, researchers have created domain-specific PLMs for improved performance on NLP tasks in the domain. In this report, we present a comprehensive examination of the clinical PLMs. More specifically, we start with a brief overview of foundational concepts of language modeling, including architectures, data sources, training methods, and more. We then introduce a list of current clinical PLMs and discuss all the models and downstream tasks in the domain. In the end, we also highlight limitations and potential future directions in the field.

1 Introduction

Text representation is a crucial task in natural language processing (NLP), forming the basis of nearly all text-related applications (Geigle et al., 2018; Liu et al., 2021b). Traditionally, to transform the input text into a vector of numerical data, one can represent the words using bag-of-words or tf-idf (term frequency-inverse document frequency) scores (Salton and Buckley, 1988; Salton, 1991) with one-hot encoding. Such methods can suffer from the *curse of dimensionality* problem as the length of vectors usually equals the size of the vocabulary, and decreased efficiency with increasing data size. Moreover, these representations fail to capture the syntactic or semantic information of the text as they only provide a statistical measure of word importance in a corpus. To overcome these issues, researchers propose word embedding techniques, e.g., Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017), to represent each word in the vocabulary with a fixed embedding vector. With the development of deep learning (LeCun et al., 2015), researchers use convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to process the text (Kim, 2014; Lai et al., 2015), with the initialization of word vectors from the aforementioned embedding methods (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017; Lu et al., 2020). This paradigm achieves significant success over a variety of downstream tasks, e.g., named-entity recognition (Sienčnik, 2015; Chiu and Nichols, 2016), text classification (Wang et al., 2016), relation classification (Zhou et al., 2016) and question answering (Xiong et al., 2017), etc. However, despite their success, word embeddings are limited in capturing polysemous words, syntactic structures, and semantic roles, hindering their full potential for use in NLP tasks (Qiu et al., 2020). For instance, the word *apple* has two different meanings in "eat an apple" and "apple computer", but it is only assigned a fixed vector according to the pre-trained word embeddings as they do not consider the contextual information during vectorization, i.e., they are *non-contextualized* or *static* embeddings.

To address the limitations of non-contextualized word embeddings, researchers have turned to the development of *contextualized* representations. With the development and emergence of the transformer architecture (Vaswani et al., 2017), considerable efforts have been put into developing transformer-based pre-trained language models (Radford et al., 2018, 2019; Devlin et al., 2019; Liu et al., 2019; Lan et al., 2019; Yang et al., 2019; Raffel et al., 2020; Lewis et al., 2020a; Brown et al., 2020; Clark et al., 2020). Essentially, the attention mechanism within the transformer allows for more GPU-based parallel computation than Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), one of the most popular and successful recurrent neural networks for text encoding, and it further facilitates large-scale pre-training and leads

to the success of the aforementioned language models. The "pre-train and fine-tune" paradigm has also been a standard approach in modern NLP for a long time. Mascio et al. present a comparative analysis on the impact of different text representation methods, i.e., BOW, traditional methods, and BERT (Devlin et al., 2019), on selected classification tasks of clinical significance (Mascio et al., 2020).

There have been plenty of pre-trained language models over the last few years, e.g., BERT (Devlin et al., 2019), GPT-1&2&3 (Radford et al., 2018, 2019; Brown et al., 2020), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), T5 (Raffel et al., 2020), BART (Lewis et al., 2020a), etc. These models roughly fall into three categories based on their different pre-training frameworks: *decoder*, *encoder*, and *encoder-decoder*. BERT (Bidirectional Encoder Representations from Transformers) drives large-scale self-supervised pre-training on extensive text corpora through the use of Masked Language Modeling (MLM). This involves masking a random subset of tokens in pre-training text and asking the model to predict the original value of the masked tokens. The self-supervised pre-training approach allows the model to learn contextualized text representations from large unannotated text corpora, such as the web, without human supervision (Wang et al., 2022). BERT also introduces Next Sentence Prediction (NSP) which aims to predict whether a given sentence follows the previous sentence or not (i.e., by `[CLS]`). Although NSP is intended to help the model understand longer-term dependencies and relationships across sentences, it is often considered unnecessary and dropped in follow-up works (Liu et al., 2019; Joshi et al., 2020; Gu et al., 2021). Unlike BERT, GPT (Generative Pre-trained Transformer) utilizes a decoder-only transformer architecture and performs an autoregressive pre-training task where they seek to predict the next token given existing ones (Radford et al., 2018). Moreover, BART (Bidirectional and Autoregressive Transformers) uses an encoder-decoder architecture and employs a denoising sequence-to-sequence pre-training task where the decoder reconstructs the original sentence from a corrupted input, and the model essentially combines bidirectional and autoregressive transformers (Lewis et al., 2020a). Generally, these models differ in their architectures, pre-training objectives, and the data they use. We will delve deeper into these differences in Section 2."

In spite of the success of these pre-trained language models on general-domain text, they struggle with domain-specific text due to the problem of *domain shift* (Ma et al., 2019). As the modern "pre-train and fine-tune" paradigm is a natural fit to domains where large-scaled unannotated textual data is available (Liu et al., 2021b), domain-specific pre-trained language models are been proposed to bridge the gap. In the biomedical and clinical domain, a variety of domain-specific PLMs have been explored and released, including BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), BlueBERT (Peng et al., 2019), ClinicalBERT (Huang et al., 2019), BioClinicalBERT (Alsentzer et al., 2019), ClinicalXLNet (Huang et al., 2020), umlsBERT (Michalopoulos et al., 2020), diseaseBERT (He et al., 2020a), ouBioBERT (Wada et al., 2020), PubMedBERT (Gu et al., 2021), SciFive (Phan et al., 2021), BioBART (Yuan et al., 2022a), ClinicalT5 (Lu et al., 2022a), etc.

Besides obtaining domain knowledge via pre-training, another line of research is knowledge infusion where domain knowledge is deliberately injected into language models to enhance their representation capability (Yao et al., 2019; Zhang et al., 2019; Kim et al., 2020; Levine et al., 2020; Wang et al., 2021b; Sun et al., 2020; He et al., 2020b; Lu et al., 2021a). One approach is to incorporate additional knowledge during pre-training. This can be achieved through an auxiliary knowledge-driven training objective. For example, KG-BERT (Yao et al., 2019) integrates factual knowledge from Wikipedia into its model through a knowledge graph completion task, while KEPLER (Wang et al., 2021b) combines a language modeling objective with a Knowledge Embedding objective for joint optimization. In the clinical domain, there is also some exploration of this direction. For instance, DiseaseBERT seeks to enhance BERT and ALBERT by incorporating disease information through additional pre-training (He et al., 2020b). DAKI (Diverse Adapters for Knowledge Integration) incorporates adapters to infuse domain knowledge of multiple sources and formats into PLMs, facilitating the integration of this knowledge in an efficient manner (Lu et al., 2021a).

It is worth noting that, though the two domains (i.e., biomedical and clinical) are relatively close and the two types of text are similar in many ways, they have some important differences. Clinical

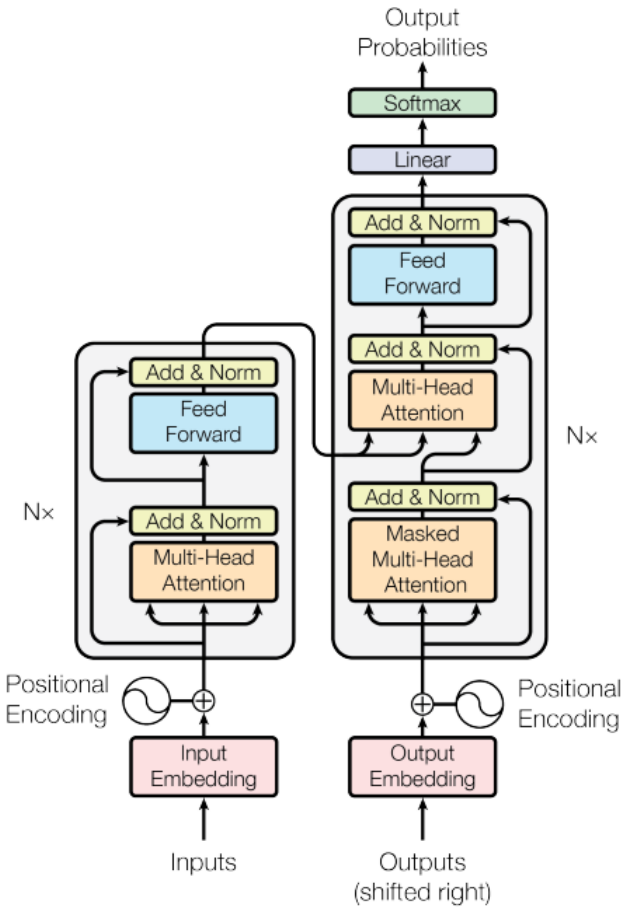

Figure 1: The Transformer model architecture (Vaswani et al., 2017).

text refers to text that is related to the practice of medicine and healthcare service, such as EHRs, physician notes, and other types of text that are commonly used in clinical settings. In contrast, biomedical text refers to text that is related to the field of biomedicine, which includes research articles, textbooks, scientific reports, and other types of text that are used in the study and advancement of biomedicine. In addition, clinical text has unique specific linguistic characteristics, such as the prevalent use of technical jargon, abbreviations, acronyms, passive verbs, and omitted subjects and verbs, which make it distinct from standard language (Smith et al., 2014). In this report, we focus on clinical PLMs and will discuss them in Section 3.

We also summarize the downstream NLP tasks in the clinical domain, as demonstrated in Section 4. For intrinsic tasks, we cover Information Extraction, Text Classification, Semantic Textual Similarity, Question Answering, Question Answering, Text Summarization, Natural Language Inference, etc. For extrinsic tasks, we discuss a bit about patients' outcomes prediction, e.g., readmission, mortality, etc, and clinical predictive tasks, e.g., diagnosis prediction. In the end, we discuss the limitations and potential future directions in Section 5.

2 Pre-trained Language Models

In this section, we first introduce the key component of modern pre-trained language models, i.e., the transformer architecture (Vaswani et al., 2017), and then discuss the most well-known general-domain PLMs in detail, e.g., BERT (Devlin et al., 2019), GPT-1&2&3 (Radford et al., 2018, 2019; Brown et al., 2020), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), T5 (Raffel et al., 2020), BART (Lewis et al., 2020a), etc.

2.1 Transformer

Recurrent neural networks (RNNs), e.g., long short-term memory networks (LSTM) (Hochreiter and Schmidhuber, 1997) and gated recurrent neural networks (GRUs), are widely adopted for sequence modeling problems such as language modeling

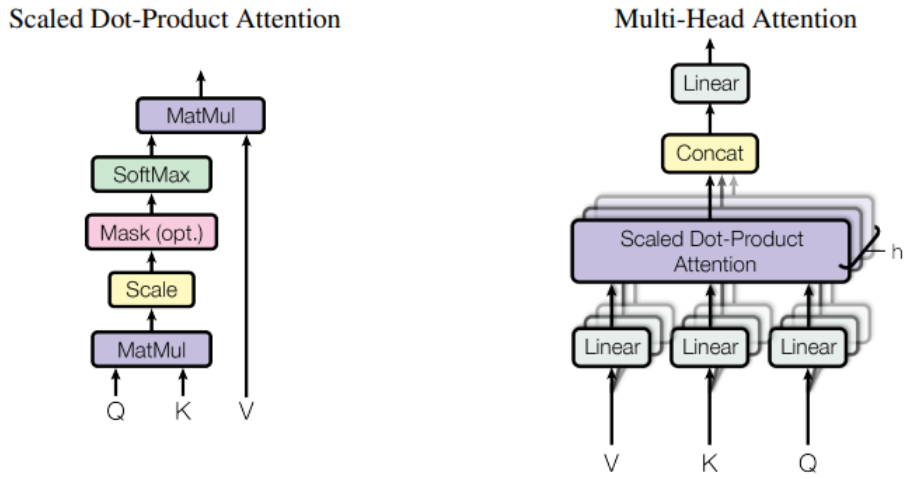

Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel (Vaswani et al., 2017).

(Bengio et al., 2000; Mikolov et al., 2010). However, the sequential nature of recurrent models often impedes parallelization within training examples, particularly with longer sequences (Vaswani et al., 2017). To overcome this limitation, Vaswani et al. introduce the Transformer, a novel transduction model architecture based solely on the *attention* mechanism, eliminating the need for recurrence (Vaswani et al., 2017). The transformer architecture allows for significantly more parallel computation and has been one of the key components of large-scale pre-trained language models.

2.1.1 Encoder and Decoder Stacks

The architecture of the transformer model is shown in Figure 1. Generally, it consists of an *encoder* and a *decoder*, both of which are stacks of transformer modules.

Encoder The encoder consists of Nx identical modules and each module has two sublayers, i.e., a multi-head self-attention layer and a position-wise fully connected feed-forward network. Within each sublayer, there is also a residual connection (He et al., 2016) and a layer normalization operation (Ba et al., 2016) that are leveraged to improve the performance and training efficiency (i.e., Add&Norm).

Decoder The decoder has a similar architecture to the encoder, except for an additional multi-head attention sublayer over the output of the encoder. The self-attention sublayer in the decoder is a bit different from that in the encoder, where future values are masked out to avoid information leakage and preserve the autoregressive property.

2.1.2 Attention

The attention mechanism is a core component in many deep learning models, especially in the field of natural language processing. It allows a model to focus its attention on specific parts of an input, such as words or phrases in a sentence when making predictions. The attention mechanism works by computing a weight for each element of the input and then using these weights to calculate a weighted sum of the elements as the output. The weights are determined by a compatibility function that measures the similarity between a query vector and key vectors associated with each element. In the Transformer architecture, attention is implemented using a combination of linear transformations and softmax activation functions. Unlike recurrent neural networks (RNNs), which use sequential computations, the linear transformations used in the Transformer's attention mechanism are relatively simple, allowing for efficient parallel computation.

Scaled Dot-Product Attention Self-attention is a mechanism used in deep learning models to capture dependencies between elements in a sequence of inputs. Essentially, it represents each input token as a weighted sum of all the token vectors in the input where the weights are computed based on the relationships between them.

In the transformer, the self-attention is implemented as "Scaled Dot-Product Attention" as shown in Figure 2. Generally, they compute the dot products of the query Q with all keys K and divide each by $\sqrt{d_k}$, and apply a softmax function

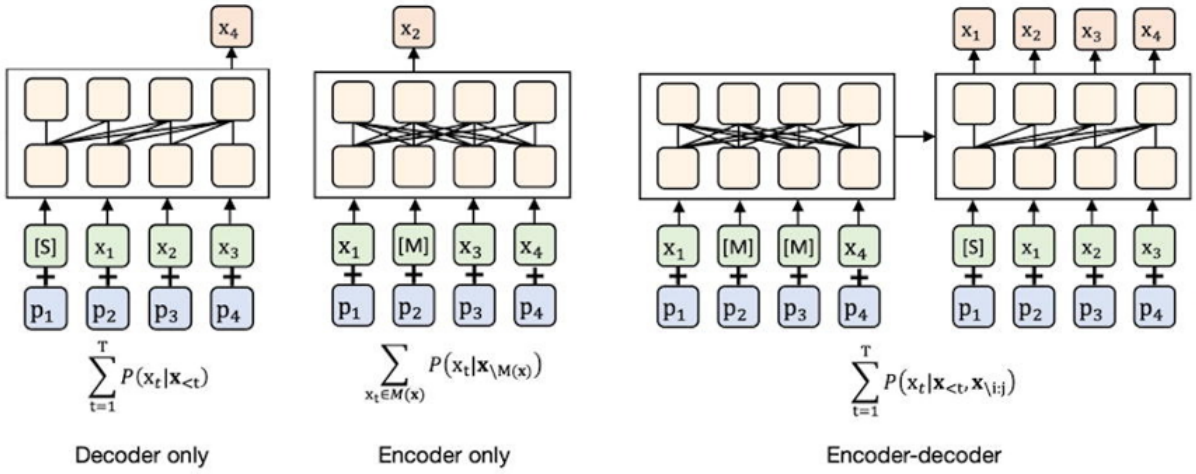

Figure 3: An illustration of existing prevalent pre-training frameworks (Wang et al., 2022).

to obtain the weights on the values V:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Multi-Head Attention Multi-head attention is a mechanism that allows a model to attend to multiple, different parts of the input sequence at once, instead of focusing on just one part as in single-head attention where the meaning of a word may largely depend on itself (Kalyan et al., 2021). In multi-head attention, the input sequence is transformed into multiple separate, parallel representations, each of which is passed through a separate attention mechanism, i.e., attention is applied multiple times in parallel. Consequently, this mechanism allows the model to capture multiple types of relationships between elements in the sequence.

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})W^O \\ \text{where head}_\text{i} &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned} \quad (2)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$,$W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$,$W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

Usage of Attention The Transformer uses multi-head attention in three different ways. The first type is the self-attention layer in the *encoder* where each position attends to all the words in the input sequence. The second type is the self-attention layer in the *decoder*. Similarly, each position attends to all positions up to that position where the future values are masked out (set to $-\infty$), i.e., masked self-attention. The third type is cross-attention within the *encoder-decoder* architecture where each position in the decoder attends to all positions in the input sequence.

2.1.3 Position-wise Feed-Forward Networks

In addition to the multi-head attention mechanism, each encoder and decoder in the Transformer architecture also includes a feed-forward neural network, as depicted in Figure 1. The feed-forward network operates in a position-independent manner, applying the same linear transformation to each element in the sequence using identical parameters. The parameters are not shared across different layers.

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b2 \quad (3)$$

2.1.4 Positional Encoding

As there are no recurrent neural networks (RNNs) that are supposed to preserve the positional information of the input sequence in the transformer, the architecture incorporates the technique *Positional Encoding* (Gehring et al., 2017) that injects a position embedding vector into individual input embeddings. This is achieved by adding a position-specific embedding vector to the embedded representation of each word. These position embedding vectors follow a learned periodic function that allows the model to determine the relative position of each word in the sequence.

2.2 Methods of PLMs

There has been a surge of interest in developing different pre-trained language models in the past few years, e.g., BERT (Devlin et al., 2019), GPT-1&2&3 (Radford et al., 2018, 2019; Brown

Model	Framework	Pre-training Method
BERT (Devlin et al., 2019)	Encoder	MLM, NSP
RoBERTa (Liu et al., 2019)	Encoder	MLM
ALBERT (Lan et al., 2019)	Encoder	MLM, SOP
XLM-R (Conneau et al., 2020)	Encoder	MLM
ELECTRA (Clark et al., 2020)	Encoder	RTD
XLNet (Yang et al., 2019)	Decoder	PLM
GPT (Radford et al., 2018)	Decoder	CLM
T5 (Raffel et al., 2020)	Encoder-Decoder	Seq2seq MLM

Table 1: Representative general-domain PLMs. Underexplored models in the clinical scenario are omitted for simplicity.

et al., 2020), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), T5 (Raffel et al., 2020), BART (Lewis et al., 2020a), etc. These models can be classified into three categories based on their pre-training frameworks: *decoder*, *encoder*, and *encoder-decoder*, as illustrated in Figure 3. In this subsection, we introduce some of the prevalent pre-training frameworks that lay the foundations of clinical PLMs and discuss their corresponding applications.

Decoder-only models (or autoregressive models) refer to models pre-trained based on the language modeling task, i.e., predicting the next token given observed ones, which also corresponds to the decoder of the transformer model. As mentioned above, GPT is a typical decoder-only pre-trained language model (Radford et al., 2018). Essentially, GPT computes the probability distribution of the next token given previous tokens, with the decoder module of the original transformer, for pre-training. The model is pre-trained on the Book Corpus dataset and demonstrates new SOTA results on several NLP benchmarks (Radford et al., 2018). GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) are the 2nd and 3rd release of GPT, which generally share the same architecture with the original version, i.e., the transformer decoder, and have 1.5 billion and 175 billion model parameters, respectively. Both GPT-2 and GPT-3 can be applied to downstream tasks without fine-tuning, demonstrating the potential of large PLMs with updated SOTA performance.

Encoder-only models (or autoencoding models) refer to models pre-trained based on the reconstruction objective of corrupted input sentences, which also corresponds to the encoder of the transformer model. Besides BERT which depends on Masked Language Modeling and Next Sentence Prediction as mentioned above, RoBERTa is another typical example of such type (Liu et al., 2019). Essentially, RoBERTa tackles some of BERT's issues and proposes the dynamic masking technique where they seek to randomly generate the mask at each epoch, as opposed to BERT's static masking strategy. RoBERTa also drops the NSP pre-training task due to its lack of impact and instead puts two consecutive full sentences together as input without asking the model to predict their consecutiveness. ALBERT (Lan et al., 2019) generally follows BERT, and it also proposes some useful tricks. Essentially, ALBERT is a light and efficient variant of BERT that differs in three aspects: (i) factorized embedding parameterization, (ii) cross-layer parameter sharing, (iii) NSP replaced by sentence ordering prediction. Empirically, the performance is better than BERT on a variety of tasks in many aspects. ELECTRA is another pre-training framework for BERT whose key innovation is Replaced Token Detection (RTD, as a replacement for MLM). The task is to simultaneously optimize a generator-discriminator architecture where the generator is trained using the MLM objective given a randomly masked sentence as input, and the discriminator (ELECTRA) aims to predict whether each token is original or generated (Clark et al., 2020). ELECTRA demonstrates efficiency and better performance than BERT/RoBERTa across multiple benchmarks.

Encoder-decoder models refer to models pre-trained based on a sequence-to-sequence objective, which also corresponds to the encoder-decoder architecture of the original transformer. As a typical example, BART (Lewis et al., 2020a) takes as input to the encoder a corrupted text with an arbitrary

noising function (Token Masking, Token Deletion, Text Infilling, Sentence Permutation, Document Rotation), and the decoder is enforced to reconstruct the original text. The model can be viewed as a combination of a bidirectional encoder (e.g., BERT) and an autoregressive decoder (e.g., GPT), and this architecture makes it better at generative tasks while keeping the bidirectional encoding capabilities. Another example is T5 (Raffel et al., 2020) which casts different NLP tasks as a text-to-text problem by assigning a specific prefix. T5 has self-supervised and supervised training. For the self-supervised pre-training, T5 takes a corrupted sentence as input and the self-supervised pre-training task is to generate the dropped-out tokens. The supervised pre-training tasks are transformed downstream tasks from the GLUE and SuperGLUE benchmarks.

Generally, there have been numerous studies on PLMs in the past few years. Some of them are CTRL (Keskar et al., 2019), Transformer-XL (Dai et al., 2019), Reformer (Kitaev et al., 2020), XLNet (Yang et al., 2019), DistilBERT (Sanh et al., 2019), ConvBERT (Jiang et al., 2020), Funnel Transformer (Dai et al., 2020), Longformer (Beltagy et al., 2020), ProphetNet (Qi et al., 2020), Switch Transformer (Fedus et al., 2021), GLaM (Du et al., 2022), Gropher (Rae et al., 2021), some multi-lingual models like mT5 (Xue et al., 2021), ERNIE (Sun et al., 2021), and so forth. As these models are rarely adopted in the clinical domain, they are not covered in this report.

3 Clinical PLMs

The rapid increase in Electronic Health Records (EHRs) and the wealth of digitized longitudinal clinical data they contain have sparked significant interest in using machine learning techniques to tackle medical challenges (Wen et al., 2019). In response to this trend, various domain-specific pre-trained language models have been developed for the clinical domain, in addition to the already existing general-domain models. In this section, we will provide a brief overview of the motivation behind developing and utilizing domain-specific pre-trained language models in the clinical field and then delve into a more in-depth examination of the different clinical PLMs available.

3.1 Motivation

In the clinical domain, the reasons for developing and utilizing domain-specific pre-trained language models are straightforward.

In general, the use of domain-specific PLMs in the clinical field is motivated by the need for improved accuracy and efficiency in language-based tasks. Training on large amounts of textual data specific to the clinical domain, such as electronic health records (EHRs) and clinical documents, enables these models to better understand and process the technical and specialized language commonly used in this field, including medical terminology and abbreviations. This can be useful for tasks such as the interpretation of EHRs, extraction of relevant information from clinical documents, generation of clinical summary reports, etc.

3.2 Data Resources

A variety of unannotated and free textual resources are used in pre-training a clinical PLM, such as clinical notes in EHRs, relevant social media posts, scientific literature, external knowledge bases, etc. We refer the readers to (Gonzalez-Hernandez et al., 2017; Kalyan and Sangeetha, 2020) for a more detailed treatment of biomedical and clinical textual corpora.

Moreover, as most domain PLMs in the biomedical and clinical domains are variants of BERT, the biggest difference among them is their pre-training data. As a result, we will cover these models, especially the less popular ones, in this subsection.

Electronic Health Records Electronic Health Records have been widely adopted by healthcare providers to electronically record patients' visits and health information in the last few years (Henry et al., 2016). There are several reasons why clinical pre-trained language models are often trained on electronic health records (EHRs). First, EHRs contain a wealth of information about patient's health histories and treatment plans, which can be valuable for language models to learn from. This information can include demographics, diagnoses, medications, laboratory test results, radiology images, and more. Second, EHRs are widely used in the healthcare industry, so clinical language models trained on EHRs may be more applicable and useful in real-world settings. Finally, since many healthcare providers use EHR systems, it is often possible to access large amounts of data from these systems for research purposes, although certain

Model	Type	Initialization	EHR
BEHRT (Li et al., 2020)	patient visits (code)	scratch	CPRD
Med-BERT (Rasmy et al., 2021)	patient visits (code)	scratch	Cerner, Truven
BRLTM (Meng et al., 2021)	patient visits (code)	scratch	private
G-BERT (Shang et al., 2019)	patient visits (code)	scratch	MIMIC-III

Table 2: Summary of EHR-based clinical PLMs.

privacy and ethical issues must be considered.

The MIMIC-III Critical Care (Medical Information Mart for Intensive Care III) Database is a large, freely-available database composed of de-identified EHR data (Johnson et al., 2016) and has been widely used for clinical NLP research (Rajkomar et al., 2018; Shorten et al., 2021; Feng et al., 2022; Lu et al., 2021c). It is also one of the most popular EHR datasets that are used to train clinical language models, which consists of the EHRs of patients in the intensive care unit (ICU) of the Beth Israel Deaconess Medical Center between 2001 and 2012.

ClinicalBERT[1] is one of the most popular domain variants which initializes from BioBERT (Lee et al., 2020) and is further pre-trained on MIMIC-III clinical notes (Alsentzer et al., 2019). Another ClinicalBERT has similar settings (Huang et al., 2019), where the authors also propose ClinicalXLNet (Huang et al., 2020), an XLNet (Yang et al., 2019) variant that is further pre-trained on MIMIC-III clinical notes. Similarly, Yang et al. propose BERT-MIMIC, ELECTRA-MIMIC, XLNET-MIMIC, RoBERTa-MIMIC, DeBERTa-MIMIC, Longformer-MIMIC based on further pre-training on MIMIC text (Yang et al., 2020). ClinicalT5 further pre-trains SciFive (Phan et al., 2021) on MIMIC notes and produces a clinical variant of T5 (Lu et al., 2022a). In general, such models mostly depend on further pre-training on unstructured clinical notes in MIMIC-III. In fact, the MIMIC database consists not only of unstructured textual data but also structured information, including different kinds of numerical features of patients, disease and procedure codes, demographics, etc.

BEHRT (Li et al., 2020) is a language model trained from scratch using EHRs, with MLM as the pre-training task. The authors use code, position, age, and segment embeddings to improve the model's performance. Med-BERT (Rasmy et al., 2021) is another language model trained from scratch with MLM and LOS (Length of Stay) as pre-training tasks. The authors use code, serialization, and visit embeddings to further improve the model's ability to handle medical data. BRLTM (Meng et al., 2021) is trained from scratch using multi-modal data with MLM. MedGPT (Kraljevic et al., 2021) is a GPT-like language model trained on patients' medical histories in the format of EHRs. Given a sequence of past medical events, MedGPT aims to predict future events.

Scientific literature Some clinical pre-trained language models are trained on scientific publications, such as research articles and medical journals because these texts can provide valuable information about current medical knowledge and practices. Scientific publications often contain detailed descriptions of medical conditions, treatments, and research findings, which can be useful for language models to learn from. Training a language model on scientific publications can also help the model to understand medical terminology and concepts more accurately and in greater depth. This can be particularly useful for tasks that involve analyzing or summarizing medical information. Finally, scientific publications may be easier to obtain than other types of clinical data, such as electronic health records (EHRs). Many scientific publications are freely available online, making it possible to create large datasets for training language models.

PubMed is a free online database that provides access to millions of scientific articles and abstracts related to medicine, biology, and life sciences. PubMed Central (PMC) is an open-access digital archive of scientific articles that contains full-text articles in the biomedical and life sciences, making it a valuable resource for researchers. PubMed abstracts (PubMed) and PubMed Central (PMC) are widely adopted for training language models in the biomedical field. (Wang et al., 2021a).

BioBERT is the first biomedical pre-trained language model which is obtained by further pre-training general BERT on biomedical literature

[1] Also known as BioClinicalBERT.

(Lee et al., 2020). Similarly, BioMedBERT is obtained by further pretraining BERT-large on the BREATHE dataset (Chakraborty et al., 2020). BlueBERT further pre-trains on the PubMed text and de-identified clinical notes from MIMIC-III (Peng et al., 2019), so as BioALBERT (Naseem et al., 2022). BioMed-RoBERTa (Gururangan et al., 2020) is obtained by further pre-training on 2.68 million full-text papers from S2ORC (Lo et al., 2020), a large corpus of academic papers spanning many academic disciplines including the biomedical domain. Unlike these models, SciBERT builds its own vocabulary and pre-trains from scratch on scientific papers from Semantic Scholar, in which 82% are from the biomedical domain and 18% are from the computer science domain (Beltagy et al., 2019). PubMedBERT is obtained by domain-specific pre-training from scratch on PubMed text (Gu et al., 2021).

Social media Clinical pre-trained language models may also be trained on social media posts, such as those from *Reddit*, *Twitter*, *AskAPatient*, *WebMD*, in order to learn about common language usage and slang in the context of healthcare. These platforms can provide a large amount of real-world data that can be used to train language models to understand how people discuss healthcare-related topics in everyday language. Training on social media posts can also provide the model with a better understanding of the context which could benefit sentiment or opinion-related tasks. However, it is important to ensure the representativeness and suitability of the data before using it for model training.

Reddit and Twitter are commonly used social media sources for training language models. Reddit is a social media platform that allows users to share news, images, and links, as well as participate in forums and discussions on a wide range of topics. Reddit is considered a valuable resource for language model training because it provides a large and diverse dataset of written content, ranging from informal conversations to in-depth discussions on a wide range of topics. Twitter is a microblogging platform that allows users to post short messages, images, and videos. Similar to Reddit, Twitter also provides a vast amount of textual data, which can help models learn to understand conversational text.

For example, BERTweet (Nguyen et al., 2020) is obtained by training on Twitter posts. COVID-twitter-BERT (Müller et al., 2020) is a natural language model to analyze COVID-19 content on Twitter. The COVID-twitter-BERT model is initialized from BERTweet and trained on tweets about COVID-19. BioRedditBERT (Basaldella et al., 2020) is initialized from BioBERT and further pre-trained on health-related Reddit posts.

External knowledge bases External medical knowledge bases can be complementary to clinical pre-trained language models, as they are often not fully exposed to structured domain knowledge, which may not be sufficiently encoded in the pre-training text. The external knowledge bases often serve more as an auxiliary training objective that works along with typical self-supervised pre-training on large amounts of textual data.

One of the most important knowledge resources is the Unified Medical Language System (UMLS)[2], which is a comprehensive and standardized terminology repository that is widely used in the field of biomedical research and healthcare (Bodenreider, 2004). It includes a wide range of medical and health-related vocabularies and terminologies, such as NCBI, MeSH, SNOMED CT, ICD-10, Gene Ontology, OMIM, and many others. The UMLS is designed to help researchers, clinicians, and other healthcare professionals communicate effectively and accurately by providing a common language and set of terms that can be used across different systems and contexts. It is maintained and updated by the National Library of Medicine (NLM) in the United States, and all vocabularies are freely available for research purposes under a corresponding license agreement.

For example, Hao et al. propose to enhance clinical BERT embedding using a joint further pre-training strategy, where they incorporate a joint loss of masked language modeling, next sentence prediction, and triplet classification on MIMIC-III notes and UMLS relations to obtain Clinical KB-BERT and Clinical KB-ALBERT (Hao et al., 2020). UmlsBERT further pre-trains ClinicalBERT (Alsentzer et al., 2019) on MIMIC-III notes with a specifically designed multi-label loss that incorporates UMLS information (Michalopoulos et al., 2020). SapBERT further pre-trains PubMedBERT (Gu et al., 2021) on UMLS synonyms under a scalable metric learning framework (Liu et al., 2021a). KeBioLM incorporates UMLS entity information by linking PubMed abstracts to the knowledge base

[2]http://umlsks.nlm.nih.gov

and adopts an entity detection/linking objective (Yuan et al., 2021). Coder injects medical knowledge from UMLS into BioBERT (Lee et al., 2020) through contrastive further training (Yuan et al., 2022b). DiseaseBERT and DiseaseALBERT are obtained by further pre-training on disease-related articles from Wikipedia (He et al., 2020a).

3.3 Pre-training Strategies

According to a recent survey on biomedical pre-trained language models, Kalyan et al. point out that existing biomedical PLMs roughly fall into the following two categories, i.e., mixed-domain pre-training, and domain-specific pre-training (Kalyan et al., 2021).

The situation in the clinical domain is quite similar. In fact, most of the aforementioned clinical/biomedical domain-specific pre-trained language models are based on the mixed-domain pre-training strategy (or continual pre-training), as pre-training on large amounts of general-domain text is proven beneficial. Essentially, the mixed-domain pre-training strategy initializes with a pre-trained model and continues the pre-training process with domain-specific data and objectives. For example, BioBERT (Lee et al., 2020) initializes from BERT (Devlin et al., 2019), ClinicalBERT (Alsentzer et al., 2019) initializes from BioBERT (Lee et al., 2020), ClinicalT5 (Lu et al., 2022a) initializes from SciFive (Phan et al., 2021), etc. This strategy demonstrates the issue of inconsistent vocabularies, which results in less representative capability of continual pre-trained models in the target domain (Gu et al., 2021). However, existing PLMs mostly use subword tokenization algorithms which effectively alleviate the issue by decomposing rare words into meaningful subwords, such as Byte-Pair Encoding (BPE) (Sennrich et al., 2016), WordPiece (Schuster and Nakajima, 2012), Unigram (Kudo, 2018), SentencePiece (Kudo and Richardson, 2018), etc.

It is important to point out that the mixed-domain pre-training approach is particularly useful when the target domain has a limited amount of text and can benefit from being pre-trained using general-domain text like Wikipedia and BookCorpus (Devlin et al., 2019) as well as related-domain text. However, this is not the case for the biomedical domain, as it has a large and growing corpus of text, with over 30 million texts in PubMed and this motivates PubMedBERT which is trained from scratch (Gu et al., 2021). Conversely, the clinical domain presents a different scenario. Due to the sensitive nature of the clinical text, such as clinical notes in EHRs, and the difficulties in obtaining such data, most clinical pre-trained language models rely on mixed-domain pre-training, such as ClinicalBERT (Alsentzer et al., 2019), ClinicalBERT (Huang et al., 2019), SciFive (Phan et al., 2021), ClinicalT5 (Lu et al., 2022a), etc.

There are also variants that are trained from scratch, such as PubMedBERT which is trained from scratch on PubMed abstracts and PMC full-text articles (Gu et al., 2021), and SciBERT which is trained from scratch on scientific papers from Semantic Scholar (Beltagy et al., 2019). Essentially, the domain-specific pre-training (training from scratch) method aims to fix the vocabulary inconsistency issue between the general domain and the biomedical domain (Kalyan et al., 2021). It is also worth noting that EHR-based language models are generally pre-trained from scratch such as BEHRT (Li et al., 2020), Med-BERT (Rasmy et al., 2021), BRLTM (Meng et al., 2021), etc., as they depend on code, demographics, visits, etc. instead of clinical narratives.

In order to gain a deeper understanding and provide a comprehensive overview of the training objectives of clinical pre-trained language models, this subsection will explore the various pre-training strategies in detail. It is worth noting that most existing clinical PLMs rely on continual pre-training, which means they would typically use similar pre-training tasks as general-domain models such as BERT (Devlin et al., 2019) but fine-tune on a large corpus of clinical data. This is done to capture the specific language and structure of the clinical domain, and improve the models' performance on downstream tasks such as named entity recognition, relation extraction, and de-identification.

In this subsection, we would cover some of the most popular pre-training tasks as well as those adopted in the aforementioned clinical PLMs.

Masked Language Modeling (MLM) This is a task where a random subset of the tokens in a sentence are replaced with a `[MASK]` token and the model is trained to predict the original token based on the context provided by the observed tokens in the sentence. Many models such as BERT (Devlin et al., 2019), BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019) use this pre-training task. As arguably one of the most pop-

Model	Type	Initialization	Corpora	Publicly Available
ClinicalBERT (Huang et al., 2019)	clinical notes	BERT	MIMIC-III	Y
ClinicalBERT (Alsentzer et al., 2019)	clinical notes	BioBERT	MIMIC-III	Y
UmlsBERT (Michalopoulos et al., 2020)	clinical notes, KG	ClinicalBERT	MIMIC-III, UMLS	Y
DiseaseBERT (He et al., 2020a)	Wiki articles	BERT	Wikipedia	Y
PubMedBERT (Gu et al., 2021)	scientific literature	scratch	PubMed, PMC	Y
BERT-MIMIC (Yang et al., 2020)	clinical notes	BERT	MIMIC-III	Y
ELECTRA-MIMIC (Yang et al., 2020)	clinical notes	ELECTRA	MIMIC-III	Y
XLNET-MIMIC (Yang et al., 2020)	clinical notes	XLNet	MIMIC-III	Y
RoBERTa-MIMIC (Yang et al., 2020)	clinical notes	RoBERTa	MIMIC-III	Y
DeBERTa-MIMIC (Yang et al., 2020)	clinical notes	DeBERTa	MIMIC-III	Y
Longformer-MIMIC (Yang et al., 2020)	clinical notes	Longformer	MIMIC-III	Y
ClinicalXLNet (Huang et al., 2020)	clinical notes	XLNet	MIMIC-III	Y
DiseaseALBERT (He et al., 2020a)	Wiki articles	ALBERT	Wikipedia	Y
BioMedBERT (Chakraborty et al., 2020)	scientific literature	BERT	BREATHE	N
BlueBERT (Peng et al., 2019)	scientific literature, clinical notes	BERT	PubMed, MIMIC-III	Y
SciBERT (Beltagy et al., 2019)	scientific literature	scratch	Semantic Scholar	Y
MedGPT (Kraljevic et al., 2021)	clinical notes	GPT	KCH, MIMIC-III	Y
BioMed-RoBERTa (Gururangan et al., 2020)	scientific literature	RoBERTa	S2ORC	Y
COVID-twitter-BERT (Müller et al., 2020)	social media posts	BERTweet	Twitter	Y
BioRedditBERT (Basaldella et al., 2020)	social media posts	BioBERT	Reddit	Y
SapBERT (Liu et al., 2021a)	KG	PubMedBERT	UMLS synonyms	Y
CODER (Yuan et al., 2022b)	KG	BioBERT	UMLS	Y
KeBioLM (Yuan et al., 2021)	KG	PubMedBERT	UMLS	Y
Clinical KB-BERT (Hao et al., 2020)	KG	BioBERT	UMLS	Y
Clinical KB-ALBERT (Hao et al., 2020)	KG	ALBERT	UMLS	Y
SciFive (Phan et al., 2021)	scientific literature	T5	PubMed, PMC	Y
BioALBERT (Naseem et al., 2022)	scientific literature	ALBERT	PubMed, PMC	Y
EhrBERT (Li et al., 2019)	clinical notes	BioBERT	private	N
RoBERTa-PubMed-MIMIC (Lewis et al., 2020b)	scientific literature, clinical notes	RoBERTa	PubMed, PMC, MIMIC-III	Y
GatorTron (Yang et al., 2022)	scientific literature, clinical notes, articles	scratch	UF Health, PubMed, Wikipedia	Y
UCSF-BERT (Sushil et al., 2022)	clinical notes	scratch	UCSF Health	N
CLIN-X-en (Lange et al., 2022)	clinical PubMed abstracts	XLM-R	PubMed	Y
CLIN-X-es (Lange et al., 2022)	clinical notes	XLM-R	Scielo archive, MeSpEn	Y
MedGTX (Park et al., 2022)	EHR	BERT	MIMIC-III	Y
Clinical-Longformer (Li et al., 2022)	clinical notes	Longformer	MIMIC-III	Y
Clinical-BigBird (Li et al., 2022)	clinical notes	BigBird	MIMIC-III	Y
BioMedLM[3]	scientific literature	GPT	PubMed, PMC	Y
DRAGON (Yasunaga et al., 2022a)	scientific literature, KG	BioLinkBERT	PubMed, UMLS	Y
Med-PaLM (Singhal et al., 2022)	instructions and exemplars	Flan-PaLM	MultiMedQA, human input	N
ClinicalT5 (Lu et al., 2022a)	clinical notes	SciFive	MIMIC-III	Y
DAKI-BERT (Lu et al., 2021a)	Wiki articles, KG	BERT	Wikipedia, UMLS	Y
DAKI-ALBERT (Lu et al., 2021a)	Wiki articles, KG	ALBERT	Wikipedia, UMLS	Y
DAKI-ClinicalBERT (Lu et al., 2021a)	Wiki articles, KG	ClinicalBERT	Wikipedia, UMLS	Y

KG = knowledge graph

Table 3: Summary of Clinical PLMs.

ular and well-explored pre-training techniques, researchers have proposed several tricks to improve its performance. For example, instead of token masking, Cui et al. propose whole word masking for Chinese BERT which demonstrates better performance (Cui et al., 2021). Besides, RoBERTa uses dynamic masking to replace BERT's static masking, where they randomly generate the mask at each epoch (Liu et al., 2019) and this trick is also applied in their domain variants, e.g., BioMed-RoBERTa (Gururangan et al., 2020). ERNIE incorporates entity-level masking and phrase-level masking which is beneficial to infuse entity knowledge into the model (Zhang et al., 2019).

Next Sentence Prediction (NSP) This task involves training the model to predict whether two sentences are contiguous or not. The objective is to learn the sentence-level context in the corpus and it's used by most of the pre-trained models derived by BERT (Devlin et al., 2019). Although NSP is intended to help the model understand longer-term dependencies and relationships across sentences, its real impact on the model has been questioned in several studies (Liu et al., 2019; Joshi et al., 2020; Gu et al., 2021), as mentioned above.

Replaced Token Detection (RTD) This is a pre-training task that is leveraged to improve robustness to word replacement and text-to-text transfer. In this task, words in a sentence are replaced with other words that have a similar meaning, and the model is trained to detect which words have been replaced. This task helps the model learn to understand the meaning of words and their relationships to other words in a sentence. One example of a model that uses RTD for pre-training is ELECTRA (Clark et al., 2020). The model uses RTD to generate masked tokens and then trains a generator model to predict the original tokens based on

the context. The generator is then fine-tuned on a downstream task and the encoder is used for the final classification. The main idea behind ELECTRA is to make the pre-training task more challenging and to reduce the risk of overfitting, by replacing some of the tokens with fake ones. It is worth noting that this task is not being widely used in the clinical domain yet. Some biomedical domain variants of ELECTRA that depend on continual pre-training naturally inherit this method, such as Bio-ELECTRA (Ozyurt, 2020), BioELECTRA (Kanakarajan et al., 2021), etc.

Sentence Order Prediction (SOP) This task aims to make the model predict the correct order of a set of sentences. Essentially, the key idea is to use two consecutive sentences from the same document as a positive sample, and to swap the two consecutive sentences to make a negative sample. This task helps the model to understand the sequential nature of language and the relationships between sentences in a document. It is worth noting that this task is motivated by the fact that NSP is often dropped by researchers due to its ineffectiveness, as mentioned above. As a replacement, ALBERT proposes SOP based on their conjecture that NSP is not very effective because it mixes both topic prediction and coherence prediction, the former of which is comparatively easy to handle which hinders the optimization of the other task (Lan et al., 2019). SOP is applied in domain variants of ALBERT, such as Clinical KB-ALBERT (Hao et al., 2020), DiseaseALBERT (He et al., 2020a), etc.

Permutation Language Modeling (PLM) This pre-training task aims to train the model to predict the correct order of a sentence given the context provided by the rest tokens of the sentence. Essentially, the input sentence is randomly permuted and the model has to reconstruct the original order by maximizing the expected log-likelihood over all possible permutations of the input. This task aims to train the model to capture bidirectional context to predict all the tokens instead of just one, which makes it more challenging than MLM. This task is applied in XLNet (Yang et al., 2019) and its domain variants ClinicalXLNet (Huang et al., 2020).

Causal language modeling (CLM) This is another name for the traditional autoregressive language modeling task, i.e., the model is trained to predict the next token given the previous tokens of the sentence. This task is typically used in autoregressive language models, e.g., GPT (Radford et al., 2018) and MedGPT (Kraljevic et al., 2021).

Sequence-to-sequence MLM This pre-training task is similar to MLM but performed in a sequence-to-sequence manner. Essentially, the input of the encoder is the corrupted sentence where random tokens are replaced by sentinel tokens, and the target is to make the decoder generate the masked tokens in an autoregressive fashion. This task is adopted in MASS (Song et al., 2019) and T5 (Raffel et al., 2020), and inherited in their domain variants SciFive (Phan et al., 2021) and ClinicalT5 (Lu et al., 2022a).

Denoising Autoencoder (DAE) This pre-training task aims to reconstruct the original sentence from a corrupted version of it, where any type of document corruption functions can be applied such as token masking, token deletion, text infilling, sentence permutation, document rotation, etc. (Lewis et al., 2020a). Essentially, the decoder reconstructs the corrupted input sentence from the output representations of the encoder (i.e., a denoising autoencoder), and the model essentially combines bidirectional and autoregressive transformers. This task is similar to some extent to seq2seq MLM in the sense that they both involve masking out or distorting a portion of the input text and then trying to predict or reconstruct that portion. This task is used in BART (Lewis et al., 2020a), BioBART (Yuan et al., 2022a).

Document Relation Prediction (DRP) This is a novel pre-training task introduced by a recent study (Yasunaga et al., 2022b). Essentially, this task aims to learn the relevance and existence of bridging concepts between documents by classifying the text segment pairs into contiguous, random, or linked. This task can be considered a variation of NSP.

Other Tasks There are other pre-training tasks that are used in specific clinical PLMs. For example, MedGTX (Park et al., 2022) claims to be the first work to propose graph-text multi-modal pre-training on EHR data. Essentially, they use a Graph Attention Networks (GAT) (Velickovic et al., 2017) based encoder to encode the structured information of an EHR, use a BERT-like model to encode the unstructured information (clinical notes), use a cross-model encoder to learn a joint representation space. Moreover, recent studies try

to encode domain knowledge into PLMs. For example, UmlsBERT (Michalopoulos et al., 2020) continually pre-trains ClinicalBERT (Alsentzer et al., 2019) on MIMIC-III notes with a specifically designed multi-label loss to inject UMLS knowledge into the model.

4 Downstream Tasks

In this section, we introduce the downstream tasks in the clinical domain, along with the corresponding datasets, that have been widely used in recent years. We first discuss the intrinsic tasks, including information extraction, text classification, word/sentence similarity, question answering, text summarization, natural language inference, etc. Then we introduce some popular extrinsic tasks, such as patient readmission prediction, mortality prediction, diagnosis prediction, and other clinical predictive tasks. It is worth noting that the distinction between intrinsic and extrinsic tasks is not always black and white, as some tasks can be considered as both intrinsic and extrinsic, e.g., text-based readmission prediction (Lu et al., 2021c).

4.1 Intrinsic Tasks

Intrinsic tasks are tasks that are primarily focused on understanding the meaning and structure of the text. These tasks are not necessarily the ones that are directly applicable to a specific domain. Examples of intrinsic tasks include, but are not limited to: information extraction, text classification, semantic textual similarity, question answering, text summarization, natural language inference, and others.

4.1.1 Information Extraction (IE)

Named Entity Recognition Named Entity Recognition (NER) is the most popular downstream NLP task in the clinical domain for the last few years, according to a recent survey (Gao et al., 2022). The task refers to identifying and classifying named entities in text into pre-defined categories such as person names, organizations, locations, medical codes, etc, and it is particularly useful for extracting structured information from unstructured text. As a specific application of NER in the clinical domain, Clinical Named Entity Recognition (CNER) aims to extract clinically relevant information, such as diseases, symptoms, treatments, medications, etc., from unstructured medical texts, e.g., clinical notes in EHRs.

A typical solution to NER is to fine-tune the PLMs to classify each token into one of the pre-defined named entity classes with a linear layer (or more advanced structures such as a LSTM layer) on top of the PLMs. This approach is often referred to as a sequence labeling task. This task has been used for evaluation for a variety of clinical and biomedical PLMs, including BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2021), etc.

Relation Extraction As is the case with NER, Relation Extraction (RE) is one of the fundamental IE tasks in the clinical scenario. Essentially, the task refers to identifying and extracting semantic relationships between two or more entities from unstructured text. And in the clinical domain, as a specific application of RE, Clinical Relation Extraction (CRE) aims to extract clinically relevant relationships between medical entities, such as causal relationships (e.g., Patient's high blood pressure caused by obesity.), symptom-disease relationships, medication-disease relationships, etc. depending on the specific task and context.

Essentially, the task is often cast as a classification problem. For example, a common approach to CRE is to fine-tune the PLMs to predict the relationships between two identified entities based on the contextual representations of the `[CLS]` token (Thillaisundaram and Togia, 2019; Su and Vijay-Shanker, 2020).

Event Extraction Event Extraction (EE) is the task that aims to identify and extract event information from text. An event can be defined as a situation or occurrence that happens at a certain point in time and has a specific set of actors, actions, and outcomes. In text, events are often described using verbs or verb phrases, and the entities involved in the event are typically described using nouns or noun phrases. For example, given a sentence "On Sunday, a protester stabbed an officer with a paper cutter.", a EE system should be able to identify an `Attack` event which consists of an event trigger `stabbed` and event arguments `Sunday`, `protester`, `officer`, `paper cutter` (Liu et al., 2020).

Similarly, Clinical Event Extraction (CEE) is a specific application of EE in the clinical domain, which aims to extract medical events from clinical text, e.g., EHRs. Medical events are occurrences or situations that happen in the medical domain, such as diagnoses, treatments, admissions, etc. For example, a CEE system should

Human Evaluation (En)

	Pair-expert	Single-expert	Single-amateur	Helpfulness
All	0.90	0.81	0.48	0.57
reddit_eli5	0.97	0.94	0.57	0.59
open_qa	0.98	0.78	0.34	0.72
wiki_csai	0.97	0.61	0.39	0.71
medical	0.97	0.97	0.50	0.23
finance	0.79	0.73	0.58	0.60

Human Evaluation (Zh)

	Pair-expert	Single-expert	Single-amateur	Helpfulness
All	0.93	0.86	0.54	0.54
open_qa	1.00	0.92	0.47	0.50
baike	0.76	0.64	0.60	0.60
nlpcc_dbqa	1.00	0.90	0.13	0.63
medicine	0.93	0.93	0.57	0.30
finance	0.86	0.84	0.84	0.75
psychology	1.00	1.00	0.60	0.67
law	1.00	0.77	0.56	0.56

Figure 4: Human evaluations of ChatGPT generated answers. (Guo et al., 2023).

extract from the sentence "Patient diagnosed with pneumonia." an event with `diagnosed` as the trigger and `Patient, pneumonia` as the arguments. Event extraction is a challenging task, especially in the clinical domain, due to the complex and private nature of this field. There have been several biomedical event extracton studies in recent years, including DeepEventMine (Trieu et al., 2020), BEESL (Ramponi et al., 2020), etc.

Entity Linking Entity Linking (EL) is a task that aims to link the entity mention in a text to its corresponding entity in a knowledge base, e.g., Wikipedia (Lu and Du, 2017; Jiang et al., 2021; Lu et al., 2022b). In the clinical domain, the task is also referred to as Medical Concept Normalization, which maps medical terms and concepts used in clinical text to a standardized terminology, such as SNOMED CT, ICD-10, or UMLS. There are some tools for this task, e.g., MetaMap (Aronson and Lang, 2010), SciSpacy (Neumann et al., 2019), etc.

Coreference Resolution Coreference Resolution is the task of identifying mentions in a text that refer to the same entity. This task is important for a wide range of NLP applications, such as information extraction, machine translation, and question answering, as it helps to understand the structure of the context and to capture the relationships between entities. In the clinical domain, coreference resolution is utilized in analyzing clinical notes, helping to support the decision-making of healthcare professionals by presenting a holistic picture of the patient and the relationships among relevant entities.

Temporal Information Extraction Temporal Information Extraction (TIE) is a task that aims to extract events or facts in the text and link them to specific times. Essentially, this task involves recognition of events and temporal expressions, recognition of temporal relations among them, and timeline construction (Leeuwenberg and Moens, 2018). TIE in the clinical domain (CTIE) aims to extract temporal information from the clinical text to understand detailed clinical observations.

De-identification : This task is to extract and mask Personal Identifiable Information (PII) from clinical notes, in order to protect patient privacy. The extracted information includes details like patient name, address, Social Security number, etc. This is a particularly important task in the clinical domain as the clinical data must comply with the Health Insurance Portability and Accountability Act (HIPAA).

4.1.2 Text Classification

Text Classification is the second most popular downstream task in the clinical domain in recent years (Gao et al., 2022). Essentially, it aims to classify input text into pre-defined categories, such as

text-based readmission prediction where they propose to predict ICU patient readmission risk using the clinical notes in EHRs (Lu et al., 2021c).

4.1.3 Semantic Textual Similarity

Semantic Textual Similarity (STS) refers to the task of predicting the degree of semantic similarity between words or sentences. The task is useful for a wide range of applications in the clinical domain, as it helps to remove redundant information that could decrease the cognitive load and enhance the clinical decision-making process (Wang et al., 2020). Typically, PLMs are used to encode the word/sentence pairs and the cosine distance is used to measure the similarity score.

4.1.4 Question Answering

Question Answering (QA) is a task that aims to extract and generate a natural language answer to a given question. Essentially, there are Extractive QA which extracts the answer from the input text, and Open/Closed Generative QA which directly generates a free-text answer to the question based on the input text. Clinical Question Answering (CQA) is a specific application of QA in the clinical domain, and it generates answers to questions related to medical information, such as diagnosis, treatment, medication, etc. CQA systems can be useful in a variety of scenarios, such as hospitals, clinics, and research institutions, to help physicians, nurses, and other healthcare professionals quickly access information and make informed decisions. CQA (or medical QA) is a challenging task as it demands comprehension of medical context, recall of appropriate medical knowledge, and reasoning with expert information (Singhal et al., 2022).

There has been a surge of interest in developing PLMs that are capable of answering questions automatically. Recently, Med-PaLM (Singhal et al., 2022) achieves state-of-the-art results on multiple medical QA benchmarks, surpassing previous models including BioMedLM, DRAGON (Yasunaga et al., 2022a), BioLinkBERT (Yasunaga et al., 2022b), Galactica (Taylor et al., 2022), PubMedBERT (Gu et al., 2021), etc. Meanwhile, ChatGPT[4] has attracted huge attention across the world and has demonstrated superior performance over a variety of tasks as shown in Figure 4, leading to a new direction for NLP research.

[4]https://chat.openai.com/chat

4.1.5 Text Summarization

Text Summarization is the task of extracting the key information of a document and generating a shorter version of it. Similar to other tasks, Clinical Text Summarization refers to the specific application of text summarization in the clinical domain, e.g., clinical notes in EHRs, etc. There are various techniques for text summarization, including extractive summarization and abstractive summarization. Extractive summarization refers to selecting and extracting the most important sentences or phrases from the original text, while abstractive summarization refers to generating a new and shorter text that summarizes the original text.

4.1.6 Natural Language Inference

Natural Language Inference (NLI) is a task that aims to predict the relationship between two sentences, i.e., a premise and a hypothesis. The goal of NLI is to classify the relationship between them as either "entailment", "contradiction", or "neutral". Clinical Natural Language Inference (CNLI) is a specific application of NLI in the clinical domain, with the goal of classifying the relationship between two pieces of clinical text. For example, given the premise "Patient has a history of hypertension and diabetes" and the hypothesis "The patient has a high risk of heart disease," the CNLI system should predict the relationship as "entailment" as the hypothesis logically follows from the premise. However, if the premise is "Patient has a history of taking aspirin for pain relief" and the hypothesis is "The patient is allergic to penicillin," the relationship should be "neutral" as there is no logical relationship between them. This task is usually cast as a ternary classification problem.

4.2 Extrinsic Tasks

Extrinsic tasks are tasks that are primarily focused on using the understanding of the text to make predictions or decisions in a specific domain. These tasks are more focused on practical or real-world problems or aspects in the specific domain. Examples of extrinsic tasks in the clinical domain include, but are not limited to: readmission prediction, mortality prediction, length of stay prediction, diagnosis prediction, and others (Lu et al., 2019, 2021b).

5 Discussion

5.1 Limitations

Insufficient Domain Expertise There have been tremendous efforts in producing stronger, faster, and larger domain-specific pre-trained language models in the clinical domain. However, most of these models depend on self-supervised pre-training over large amounts of textual data, e.g., ChatGPT uses 175 billion parameters and Med-PaLM has 540 billion parameters (Singhal et al., 2022). Recently, ChatGPT has attracted attention all over the world as the model shows remarkable performance on different kinds of NLP-related tasks across multiple domains, including the biomedical and clinical fields. However, the model is still considered "unhelpful" for medicine as judged by human experts as against other domains (e.g., as shown in Figure 4), revealing that the seemingly almighty model lacks an in-depth understanding of domain knowledge (Guo et al., 2023). In fact, there has been a surge of interest in proposing novel methods to inject domain knowledge into existing PLMs (He et al., 2020a; Lu et al., 2021a; Michalopoulos et al., 2020). Nevertheless, these works mostly focus on empirical improvement over different benchmarks without providing an in-depth and clear explanation of how the infused knowledge actually affects the model inference, which could limit their impact.

Data Scarcity Another limitation of the clinical PLMs is the limited availability of their pre-training data. Essentially, most of the aforementioned clinical PLMs depend on clinical notes, e.g., the MIMIC database (Alsentzer et al., 2019; **?**), which is relatively small in size and does not support the training of larger models (Johnson et al., 2016). This scarcity of data can negatively impact the performance of the models and limit their ability to generalize to real-world scenarios.

Interpretability Despite the impressive performance of clinical PLMs, their lack of interpretability remains an issue, as it can limit the trust placed in the models and their ability to be used in real-world clinical settings.

Privacy, Security and Ethical considerations Clinical PLMs often work with sensitive patient information, making privacy and security a major concern. There is a need to ensure that patient data is protected and kept confidential, which can be challenging in the context of Clinical NLP. The use of clinical PLMs also raises important ethical considerations, such as the potential for algorithmic bias and discrimination, the responsibility for the outputs of the models, and the potential impact on patient care and outcomes.

5.2 Future Directions

One promising avenue of future research is to investigate novel pre-training methods that incorporate large amounts of domain knowledge from knowledge bases and limited amounts of clinical notes. The "big knowledge, small data" approach may provide a solution to the challenges of insufficient domain expertise and data scarcity that are faced by current clinical PLMs.

Another important direction is to delve deeper into the interpretability issue of clinical PLMs and their applications. Understanding the thought process and reasoning behind physician diagnoses can provide valuable insights into the use of clinical PLMs. Furthermore, exploring the impact of diverse sources of domain knowledge on model inference can help to better understand how to effectively incorporate knowledge into clinical PLMs. This can lead to improved model performance and increased trust in applying machine learning techniques in the clinical setting.

6 Conclusion

In this report, we provide a comprehensive overview of pre-trained language models in the clinical domain. We begin by introducing the key concepts of pre-training methods, model architectures, pre-training data, and other relevant information. Next, we present an extensive list of current clinical PLMs, highlighting their key features and characteristics. Finally, we delve into the limitations of current clinical PLMs, including issues related to the lack of domain knowledge and data scarcity. Finally, we conclude by exploring future directions for Clinical NLP, including the development of novel pre-training methods and a deeper understanding of model interpretability and its applications in the clinical setting.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clin-*

ical Natural Language Processing Workshop (ClinicalNLP), pages 72–78.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611.

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

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. Biomedbert: A pre-trained biomedical language model for qa and ir. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. 2020. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. *Advances in neural information processing systems*, 33:4271–4282.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL-HLT)*, pages 4171–4186.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

Jean Feng, Rachael V Phillips, Ivana Malenica, Andrew Bishara, Alan E Hubbard, Leo A Celi, and Romain Pirracchio. 2022. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of ai algorithms in healthcare. *npj Digital Medicine*, 5(1):1–9.

Yanjun Gao, Dmitriy Dligach, Leslie Christensen, Samuel Tesch, Ryan Laffin, Dongfang Xu, Timothy Miller, Ozlem Uzuner, Matthew M Churpek, and Majid Afshar. 2022. A scoping review of publicly available language tasks in clinical natural language processing. *Journal of the American Medical Informatics Association*, 29(10):1797–1806.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Chase Geigle, Qiaozhu Mei, and ChengXiang Zhai. 2018. Feature engineering for text data. In *Feature engineering for machine learning and data analytics*, pages 15–54. CRC Press.

Graciela Gonzalez-Hernandez, Abeed Sarker, Karen O'Connor, and Guergana Savova. 2017. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, 26(01):214–227.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Boran Hao, Henghui Zhu, and Ioannis C Paschalidis. 2020. Enhancing clinical bert embedding using a biomedical knowledge base. In *28th International Conference on Computational Linguistics (COLING 2020)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020a. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020b. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614, Online. Association for Computational Linguistics.

JaWanna Henry, Yuriy Pylypchuk, Talisha Searcy, Vaishali Patel, et al. 2016. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015. *ONC data brief*, 35(35):2008–2015.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward Moseley, Chih-Ying Deng, Naomi George, and Charolotta Lindvall. 2020. Clinical XLNet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 94–100, Online. Association for Computational Linguistics.

Hang Jiang, Sairam Gurajada, Qiuhao Lu, Sumit Neelam, Lucian Popa, Prithviraj Sen, Yunyao Li, and Alexander Gray. 2021. LNN-EL: A neuro-symbolic approach to short-text entity linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 775–787, Online. Association for Computational Linguistics.

Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. 2020. Convbert: Improving bert with span-based dynamic convolution. *Advances in Neural Information Processing Systems*, 33:12837–12848.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, page 103982.

Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. 2020. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics*, 101:103323.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Bosung Kim, Taesuk Hong, Youngjoong Ko, and Jungyun Seo. 2020. Multi-task learning for knowledge graph completion with pre-trained language models. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 1737–1743.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. 2021. Medgpt: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2267–2273.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2022. Clin-x: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain. *Bioinformatics*, 38(12):3267–3274.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

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

Artuur Leeuwenberg and Marie Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1237–1246.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. Sensebert: Driving some sense into bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4656–4667.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020b. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157.

Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, Hong Yu, et al. 2019. Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study. *JMIR medical informatics*, 7(3):e14830.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 1641–1651, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

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

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Qiuhao Lu, Nisansa de Silva, Dejing Dou, Thien Huu Nguyen, Prithviraj Sen, Berthold Reinwald, and Yunyao Li. 2020. Exploiting node content for multiview graph convolutional network and adversarial regularization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 545–555, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qiuhao Lu, Nisansa De Silva, Sabin Kafle, Jiazhen Cao, Dejing Dou, Thien Huu Nguyen, Prithviraj Sen, Brent Hailpern, Berthold Reinwald, and Yunyao Li. 2019. Learning electronic health records through hyperbolic embedding of medical ontologies. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 338–346.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022a. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qiuhao Lu, Dejing Dou, and Thien Huu Nguyen. 2021a. Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3855–3865, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qiuhao Lu, Dejing Dou, and Thien Huu Nguyen. 2021b. Textual data augmentation for patient outcomes prediction. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2817–2821. IEEE.

Qiuhao Lu and Youtian Du. 2017. Wikipedia-based entity semantifying in open information extraction. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 765–770. IEEE.

Qiuhao Lu, Sairam Gurajada, Prithviraj Sen, Lucian Popa, Dejing Dou, and Thien Nguyen. 2022b. Cross-lingual short-text entity linking: Generating features for neuro-symbolic methods. In *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, pages 8–14, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Qiuhao Lu, Thien Huu Nguyen, and Dejing Dou. 2021c. Predicting patient readmission risk from medical text via knowledge graph enhanced multiview graph convolution. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1990–1994.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with bert-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo)*, pages 76–83.

Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020. Comparative analysis of text classification approaches in electronic health records. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 86–94, Online. Association for Computational Linguistics.

Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. 2021. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3121–3129.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alex Wong. 2020. Umls-bert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. *arXiv preprint arXiv:2010.10391*.

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

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Usman Naseem, Adam G Dunn, Matloob Khushi, and Jinman Kim. 2022. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *BMC bioinformatics*, 23(1):1–15.

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*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Ibrahim Burak Ozyurt. 2020. On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 104–112, Online. Association for Computational Linguistics.

Sungjin Park, Seongsu Bae, Jiho Kim, Tackeun Kim, and Edward Choi. 2022. Graph-text multi-modal pre-training for medical representation learning. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 261–281. PMLR.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):1–10.

Alan Ramponi, Rob van der Goot, Rosario Lombardo, and Barbara Plank. 2020. Biomedical event extraction as sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5357–5367, Online. Association for Computational Linguistics.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Gerard Salton. 1991. Developments in automatic text retrieval. *science*, 253(5023):974–980.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1):1–34.

Scharolta Katharina Sienčnik. 2015. Adapting word2vec to named entity recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 239–243.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Kelly Smith, Beata Megyesi, Sumithra Velupillai, and Maria Kvist. 2014. Professional language in swedish clinical text: Linguistic characterization and comparative studies. *Nordic Journal of Linguistics*, 37(2):297–323.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936. PMLR.

Peng Su and K Vijay-Shanker. 2020. Investigation of bert model on biomedical relation extraction based on revised fine-tuning mechanism. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2522–2529. IEEE.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. CoLAKE: Contextualized language and knowledge embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3660–3670, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Madhumita Sushil, Dana Ludwig, Atul J Butte, and Vivek A Rudrapatna. 2022. Developing a general-purpose clinical language inference model from a large corpus of clinical notes. *arXiv preprint arXiv:2210.06566*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Ashok Thillaisundaram and Theodosia Togia. 2019. Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 84–89, Hong Kong, China. Association for Computational Linguistics.

Hai-Long Trieu, Thy Thy Tran, Khoa NA Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. Deepeventmine: end-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19):4910–4917.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6000–6010.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *stat*, 1050:20.

Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2020. Pre-training technique to localize medical bert and enhance biomedical bert. *arXiv preprint arXiv:2005.07202*.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, et al. 2021a. Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-trained language models and their applications. *Engineering*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics (TACL)*, 9:176–194.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54:57–72.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Andrew Wen, Sunyang Fu, Sungrim Moon, Mohamed El Wazir, Andrew Rosenbaum, Vinod C Kaggal, Sijia Liu, Sunghwan Sohn, Hongfang Liu, and Jungwei Fan. 2019. Desiderata for delivering nlp to accelerate healthcare ai advancement and a mayo clinic nlp-as-a-service implementation. *NPJ digital medicine*, 2(1):1–7.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dcn+: Mixed objective and deep residual coattention for question answering. *arXiv preprint arXiv:1711.00106*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy Liang, and Jure Leskovec. 2022a. Deep bidirectional language-knowledge graph pretraining. *arXiv preprint arXiv:2210.09338*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022b. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022a. Biobart: Pretraining and evaluation of a biomedical generative language model. *arXiv preprint arXiv:2204.03905*.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190, Online. Association for Computational Linguistics.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022b. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, 126:103983.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1441–1451.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.