

# A Survey on Biomedical Text Summarization with Pre-trained Language Models

Qianqian Xie, Zheheng Luo, Benyou Wang, Sophia Ananiadou

**Abstract**—The exponential growth of biomedical texts such as biomedical literature and electronic health records (EHRs), provides a big challenge for clinicians and researchers to access clinical information efficiently. To address the problem, biomedical text summarization has been proposed to support clinical information retrieval and management, aiming at generating concise summaries that distill key information from single or multiple biomedical documents. In recent years, pre-trained language models (PLMs) have been the de facto standard of various natural language processing tasks in the general domain. Most recently, PLMs have been further investigated in the biomedical field and brought new insights into the biomedical text summarization task. In this paper, we systematically summarize recent advances that explore PLMs for biomedical text summarization, to help understand recent progress, challenges, and future directions. We categorize PLMs-based approaches according to how they utilize PLMs and what PLMs they use. We then review available datasets, recent approaches and evaluation metrics of the task. We finally discuss existing challenges and promising future directions. To facilitate the research community, we line up open resources including available datasets, recent approaches, codes, evaluation metrics, and the leaderboard in a public project: <https://github.com/KenZLuo/Biomedical-Text-Summarization-Survey/tree/master>.

**Index Terms**—Biomedical texts, text summarization, pre-trained language models.

## 1 INTRODUCTION

WITH the rapidly increasing of unstructured clinical information, such as biomedical literature [1] and clinical notes [2], there is a big challenge for researchers and clinicians to access the required information effectively. To address it, the text summarization technique [3] has been explored in the biomedical domain to help users seek information more efficiently. Biomedical text summarization [4] aims to shorten single or multiple long biomedical documents into a condensed summary that keeps the most important semantic information. It saves much time and human effort for users since they can grasp the main idea of long biomedical documents by only reading the summary quickly. It can be applied in various real applications including but not limited to aiding evidence-based medicine [5], clinical information management [6], and clinical decision support [7].

In recent years, pre-trained language models (PLMs) [8] that are applied as the paradigm of various natural language processing tasks, have been introduced into biomedical text summarization [9], [10]. Compared with existing methods, such as graph-based ranking methods [11], traditional machine learning methods [12] and deep learning methods [13], self-supervised pre-training makes PLMs memorize common sense and lexical knowledge inherited in the training texts [14], which can be transferred to improve NLP

tasks via fine-tuning. Without manually annotated data, PLMs can greatly boost the performance of various NLP tasks via knowledge transfer where large-scale unlabeled data is available, such as text summarization in the biomedical domain. In these methods, general-domain PLMs such as BERT [8] or domain-specific PLMs such as BioBERT [15] is employed as the backbone model for encoding input texts. They are further fine-tuned with the specific loss of the biomedical text summarization task on the biomedical unstructured dataset. It allows the semantic knowledge captured in PLMs to be transferred to the biomedical text summarization task, resulting in more conclusive and informative summaries.

Despite the fact that there were previous surveys for traditional machine learning and deep learning techniques on biomedical text summarization [4], [16], [17], [18], there were no efforts summarizing and tracking the recent development of PLMs on biomedical text summarization task. To fill the gap, this paper surveys recent work that utilizes PLMs for the biomedical text summarization task. We systematically review benchmark datasets, recent approaches, and evaluation methods of the task. We categorize and discuss existing methods according to how they use PLMs (feature-based, fine-tuning based, and adaption+fine-tuning based), and what PLMs they used (encoder-based PLMs, decoder-based PLMs, and encoder-decoder based PLMs). We hope this paper can be a timely survey for researchers in the research community to quickly track recent progress, challenges, and promising future directions. The main contributions of this survey are:

- We propose a comprehensive review of biomedical text summarization with pre-trained language models. To the best of our knowledge, this is the first

• Qianqian Xie, Zheheng Luo, and Sophia Ananiadou are with the Department of Computer Science, University of Manchester, Manchester, United Kingdom.  
E-mail: [qianqian.xie@manchester.ac.uk](mailto:qianqian.xie@manchester.ac.uk); [zheheng.luo@postgrad.manchester.ac.uk](mailto:zheheng.luo@postgrad.manchester.ac.uk); [Sophia.Ananiadou@manchester.ac.uk](mailto:Sophia.Ananiadou@manchester.ac.uk)  
Benyou Wang is with Department of Computer Science, The Chinese University of Hong Kong, Shengzhen, China.  
E-mail: [wangbenyou@cuhk.edu.cn](mailto:wangbenyou@cuhk.edu.cn)  
[Sophia.Ananiadou@manchester.ac.uk](mailto:Sophia.Ananiadou@manchester.ac.uk) is the corresponding author

review that surveys recent PLMs-based methods.

- We categorize and discuss recent approaches, benchmark datasets, and evaluation methods thoroughly.
- We discuss challenges of existing approaches and outlook promising future directions.

**Compared with existing surveys** Afantenos et al [16] was the earliest survey that summarized traditional natural language processing and machine learning methods for medical document summarization. Mishra et al [4] reviewed text summarization methods for biomedical literature and electronic health records (EHRs), between January 2000 and October 2013. Pivovarov et al [17] examined automated summarization methods for electronic health records. Most methods summarized in these surveys are traditional machine learning methods based on feature engineering. With the prosperity of deep learning since 2014, deep neural networks became the mainstream method for biomedical text summarization. Recently, Wang et al [18] investigated deep learning-based text summarization approaches for both biomedical literature and EHRs between January 2013 to April 2021. Nevertheless, researches with PLMs for biomedical text summarization were not included in this survey. Although there were previous efforts summarizing PLMs-based methods for biomedical natural language processing [19], they only briefly introduced PLMs for the biomedical text summarization task as one of the various tasks. Compared with them, we provide a more comprehensive and focused overview for PLMs on biomedical text summarization including benchmark datasets, evaluation metrics, and limitations et al.

**Paper collection** We collect representative works since 2018 that are published in conferences and journals of computer science and biomedical science such as ACL, EMNLP, COLING, NAACL, AACL, Bioinformatics, BioNLP, JAMIA, AMIA, NPJ digital medicine et al. We use PubMed and google scholar as the search engine and the database. We search with keywords including "biomedical summarization", "medical summarization", "clinical summarization", "medical dialogue summarization" et al.

**Organization of the paper** We will first introduce the background of biomedical text summarization and pre-trained language models in the Section 2. Then Section 3 will describe benchmark datasets. Representative PLMs based methods will be categorized and discussed in Section 4. We introduce evaluation methods in Section 5. We next discuss limitations and future directions in Section 6. Finally, we make a conclusion in Section 7. Figure 1 shows the proposed overview of biomedical text summarization with pre-trained language models.

## 2 BACKGROUND

In this section, we first review biomedical text summarization and pre-trained language models, which are two essential concepts used in this survey. The overview of the background section is shown in Figure 2.

### 2.1 Biomedical Text Summarization

Biomedical text summarization aims to shorten single or multiple biomedical documents into a condensed summary

that preserves the most important information from the original text. In general, automated summarization approaches are divided into extractive summarization methods [20] and abstractive summarization methods [21] according to the output of summaries as shown in Figure 3. Extractive methods select key sentences from original documents and concatenate them into a summary, while abstractive methods generate new sentences as the summary based on the original documents. Compared with extractive summarization, abstractive summarization is more challenging. It is difficult for automated abstractive methods to generate factually consistent summaries since it involves generating informative sentences from the large vocabulary, lexical and syntactic adjustment, and paraphrasing. Formally, let's assume  $C$  as a biomedical corpus with  $D$  documents,  $d \in C$  is a document consisting of  $m$  sentences:  $d = \{s_1, \dots, s_m\}$ . We also assume the gold summary of the document  $d$  as  $t_d$ . For biomedical scientific papers, abstracts of papers are generally deemed as their gold summaries.

Automatic biomedical summarization methods are largely facilitated and inspired by automatic methods in the general domain. The earliest methods are traditional machine learning methods such as Naive-Bayes classifier [22], and graph-based ranking methods such as TextRank [11]. With the prosperity of deep learning since 2014, neural network methods have been the mainstream method for both extractive and abstractive summarization of biomedical texts. Neural extractive methods [20] formulate the extractive task as the binary classification problem that predicts labels (1 or 0) of sentences in original documents to select sentences. As for neural abstractive methods [23], they model the abstractive task as the text generation problem that generates new sentences based on the sequence-to-sequence [24] framework.

**Extractive summarization** For document  $d$ , extractive summarization methods aim to select a subset of  $o$  sentences from  $d$ ,  $o \ll m$ . Neural extractive methods can be classified into unsupervised methods and supervised methods. The unsupervised methods model the extractive task into the sentence ranking problem. They generate sentence representations based on word embeddings, and use the unsupervised ranking method to select important sentences based on their representations. For supervised methods, since most corpus adopt human-written abstractive summaries as the gold summaries whose sentences are not in the original documents, they are first required to generate binary labels for sentences according to the gold summaries, to train the extractive models. To this end, they generally adopt unsupervised sentence selection methods such as the greedy search algorithm [20] to generate the oracle summary for each document with sentences that are most semantically similar to the gold summary. Therefore, sentences that are included in the oracle summary are labeled with 1, while the remaining sentences are labeled with 0.

Most supervised neural extractive methods consist of the neural network-based encoder and classifier. A neural network-based encoder is used to capture the contextual information of input documents and generate vector representations of sentences. The classifier is to predict labels of sentences according to their vector representations. The objective is to maximize the log-likelihood of the observed

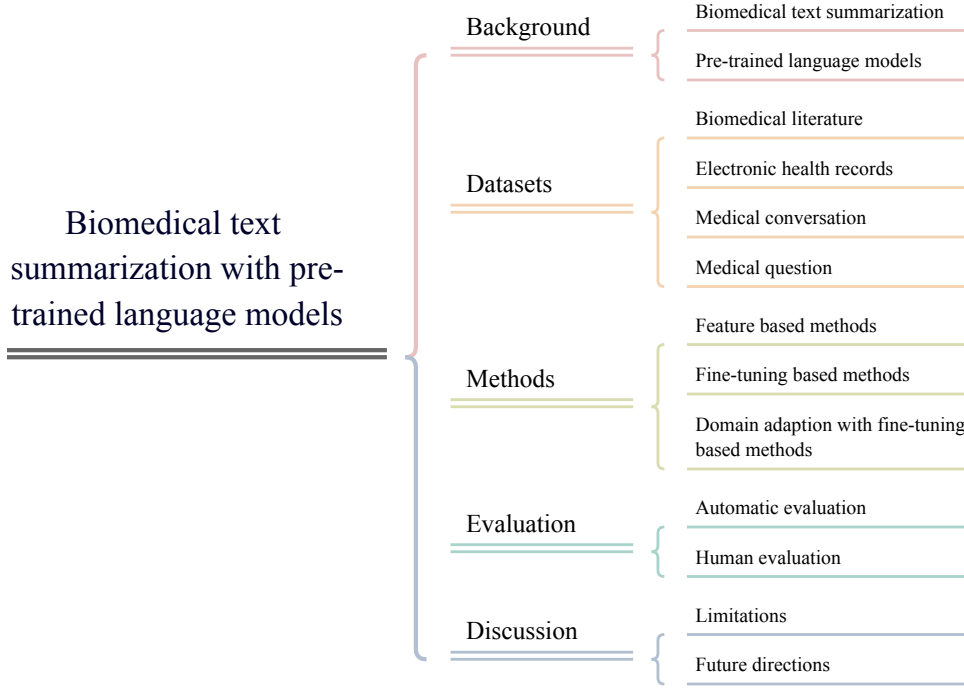


Fig. 1. Overview of biomedical text summarization with pre-trained language models.

labels of sentences:

$$\log p(y|C; \theta) = \sum_{d \in C} \sum_{i=1}^m \log p(y_i^d | d; \theta) \quad (1)$$

where  $y_i^d$  is the ground truth label of sentence  $s_i$  in document  $d$ ,  $\theta$  is the parameter set of the model.

**Abstractive summarization** Neural abstractive methods build the abstractive task as the sequence-to-sequence learning problem. Most of them utilize the encoder-decoder framework [24], which consists of the neural network-based encoder and decoder. Similar to extractive methods, the encoder is used to yield vector representations of input documents. The decoder is to generate the target summary sequentially with representations from the encoder. The model is optimized via the objective to maximize the log-likelihood of target words in the gold summary.

$$\log p(t|C; \theta) = \sum_{d \in C} \sum_{i=1}^n \log p(t_i^d | d; \theta) \quad (2)$$

where  $t_i^d$  is the  $i$ -th word in the gold summary  $t^d$  of the document  $d$ ,  $n \ll m$ .

Most recently, PLMs have become the new paradigm of biomedical summarization. PLMs-based methods have a similar framework to neural methods for extractive and abstractive tasks, while PLMs are more powerful than neural networks in encoding biomedical texts. We will next introduce the pre-trained language models.

## 2.2 Pre-trained Language Models

Language model pre-training [25], [26] has long been an active research area with the aim of learning low-dimensional vector representations from natural language, which are applicable and generalizable for downstream

tasks. The earliest unidirectional neural language models such as word2vec [26] and glove [27], learn meaningful word embeddings via estimating the probability of the next word with the sequence of history words. The bidirectional language models such as ELMo [28] are then proposed to further consider the bidirectional context of words. Bidirectional Encoder Representations from Transformers (BERT) [8] is the breakthrough work that advances the state-of-art of various NLP tasks. BERT and its variants generally consist of two steps: pre-training and fine-tuning. It proposes to first pre-train the deep models based on basic neural network structure such as transformer [29] on the large scale of unlabeled data with a self-supervised learning task. Then the pre-trained parameters of deep models and task-specific parameters are fine-tuned on labeled data with downstream tasks. We will further illustrate the core components of PLMs along with pre-training and fine-tuning. For more details of PLMs, one can check the review [30].

**Model architecture** The early language models such as ELMo and its predecessors [28], [31], generally utilize Bi-LSTM [32] as the backbone network structure, to capture bi-directional contextual information of texts. However, Bi-LSTM has the limitation of parallelization and sequential computation with the growth of sequence length. One breakthrough work is Transformer [29], which proposes the self-attention-based neural network model architecture. It is able to parallel computation and model long-range dependencies of sequences efficiently. The Transformer follows the encoder-decoder architecture with stacked multi-head self-attention and a point-wise fully connected feed-forward network. After that, nearly all pre-trained language models utilize the Transformer architecture. To learn better representations, they usually have deep network architecture. For example, the base model of BERT has 12 Transformer layers

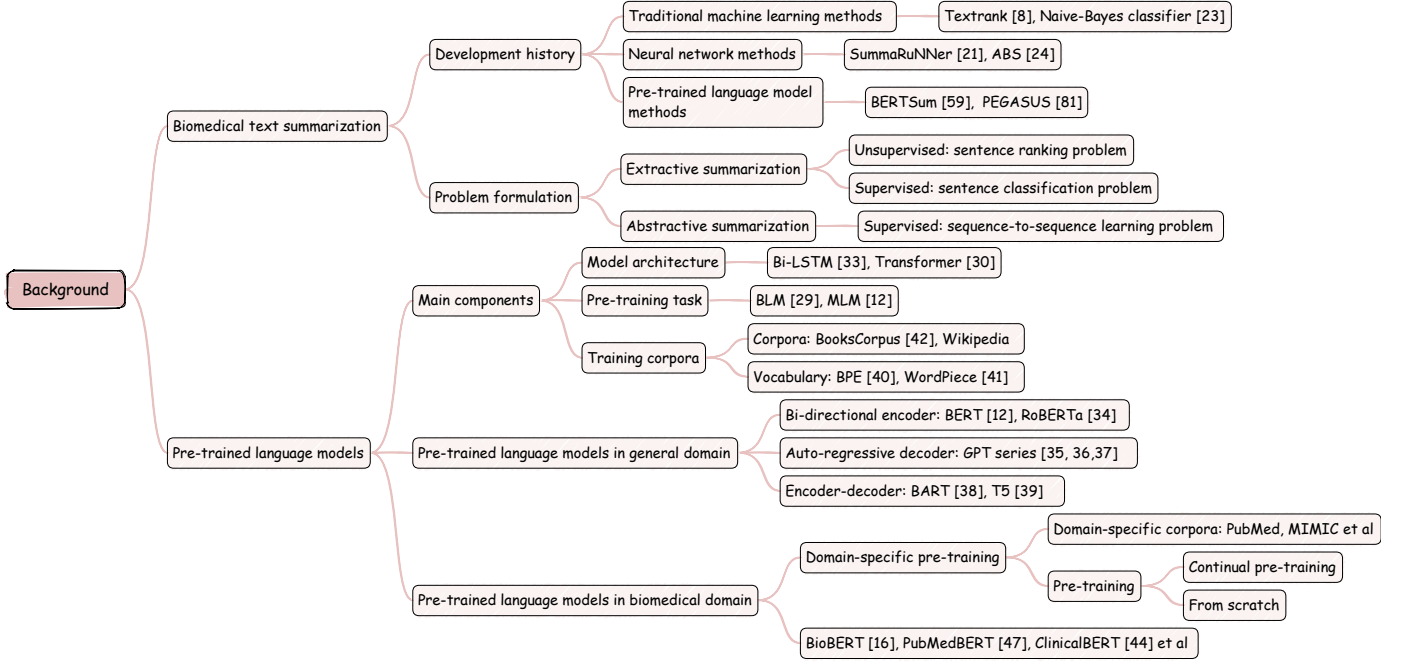


Fig. 2. Overview of background.

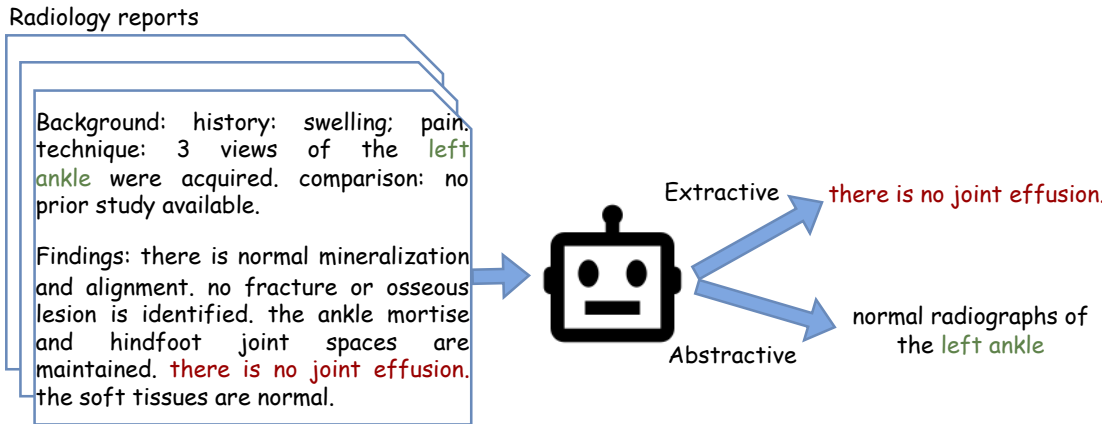


Fig. 3. The example of extractive and abstractive biomedical text summarization.

with hidden size 768 and 12 self-attention heads.

According to different model architectures, existing PLMs can be categorized into three types: bi-directional encoder language models, auto-regressive decoder language models, and encoder-decoder language models, as shown in Figure 4. The bi-directional language models, i.e., BERT and its variants such as Roberta [33], use Transformer as the bi-directional encoder without using the decoder structure in Transformer. The auto-regressive language models such as GPT series [34], [35], [36], only pre-train auto-regressive decoders based on Transformer architecture. Different from these two methods, encoder-decoder language models such as BART [37] pre-train the full encoder-decoder Transformer architecture.

**Training corpora** Most PLMs use the corpora in the general domain such as BooksCorpus [38] and Wikipedia. To address the out-of-vocabulary words, they split words into sub-words to formulate the vocabulary via the Byte-

Pair Encoding (BPE) [39] or WordPiece [40] methods.

**Pre-training** As the first step of PLMs, the pre-training task on a large scale of unlabeled data is the key for language models to learn useful representations and parameters, which can be fine-tuned to downstream tasks. The pre-training task of most previous language models follows the unidirectional language model [25]. It aims to maximize the log-likelihood of words conditionally on history words:

$$\mathcal{L}_{lm} = - \sum_{t=1}^T \log p(x_t | x_1, x_2, \dots, x_{t-1}) \quad (3)$$

where  $X = \{x_1, \dots, x_T\}$  is a given text sequence with  $T$  words. The bidirectional language model is further proposed to capture contextual information of text from both directions. It combines both the left-to-right language model

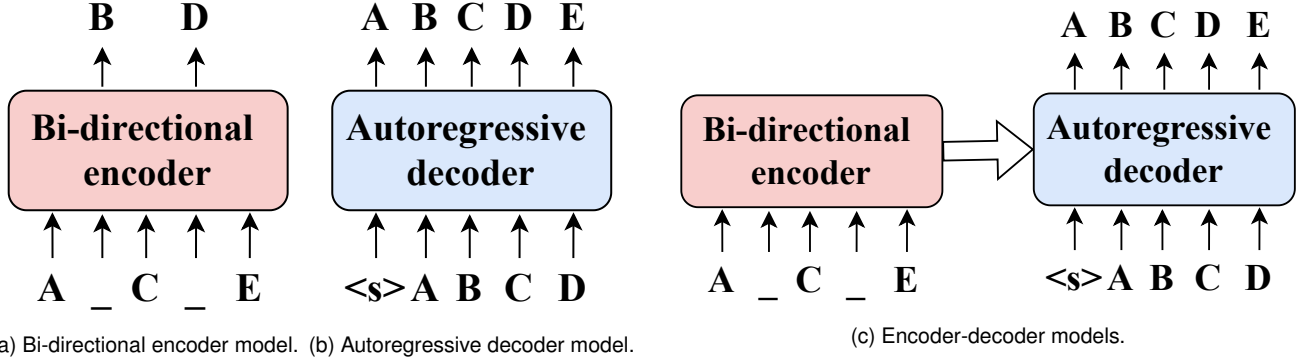


Fig. 4. Comparison of different language models. Bi-directional encoder language models use only the encoder portion of Transformer and predict masked tokens independently with the masked input sequence. They are hard to be used for generation tasks. Different from the bi-directional encoder language models, autoregressive decoder language models are uni-directional and only use the decoder portion of the Transformer. They predict tokens auto-regressively and therefore can be used for generation tasks. Encoder-decoder language models use the full encoder-decoder architecture in Transformer. They are more flexible since there is no need to align the input tokens of the encoder and output tokens in the decoder.

and right-to-left language model:

$$\mathcal{L}_{blm} = - \sum_{t=1}^T (\log p(x_t | x_1, x_2, \dots, x_{t-1}) + \log p(x_t | x_{t+1}, x_{t+2}, \dots, x_T)) \quad (4)$$

Different from the bidirectional language model, PLMs such as BERT utilize the masked language model (MLM), which allows bi-directional self-supervised pre-training more efficiently. It randomly selects 15% tokens of the input text to predict, in which 80% of them are replaced with the special token "[MASK]", 10% of them are replaced with other words in the vocabulary. The objective is to maximize the log-likelihood of ground-truth words in the selected positions with masked text sequence:

$$\mathcal{L}_{mlm} = - \sum_{\hat{x} \in m(x)} \log p(\hat{x} | X_m) \quad (5)$$

where  $X_m$  is the masked text sequence and  $m(x)$  is the set of masked words. Pre-training with MLM guides the language model to fully capture the contextual information embedded in the token sequence and generate more expressive representations on different levels such as tokens and sentences.

**Fine-tuning** Although self-supervised pre-training on a large-scale corpus allows language models to memorize common sense and linguistic knowledge in pre-trained parameters and contextual representations, it is still essential to adapt the model and generated representations to downstream tasks via fine-tuning with task-specific objectives and datasets. According to downstream tasks, task-specific inputs are firstly fed into pre-trained language models to yield contextual representations. Different tasks usually are formulated into different problems such as classification, regression, and generation. Therefore, it requires choosing contextual representations on different levels and different task-specific layers stacked on top of language models. For example, for the extractive summarization task, previous researches generally append an extra classification layer to predict the labels of sentences based on the sentence representations from PLMs. By optimizing the model with

the classification loss, the parameters of PLMs and task-specific parameters are refined. The general linguistic and semantic knowledge in the pre-trained language models are transferred into task-specific representations via fine-tuning, which have shown great performance in various tasks and become the paradigm of these tasks.

### 2.3 Biomedical Language Models

Inspired by the great success of PLMs on NLP tasks, much attention has been devoted to applying PLMs to tasks in the biomedical domain including biomedical text summarization. However, most advanced pre-trained language models, including BERT, variants of BERT, GPT3 [36], T5 [41] et al, are pre-trained on texts of the general domain such as BooksCorpus [38] and Wikipedia. It is challenging to directly apply these models to biomedical texts. The greatest issue is the terminologies and compound words in biomedical texts, most of which have never been mentioned in the general domain texts. These domain-specific tokens uncovered by the vocabulary of PLMs nevertheless embed the salient information which is fundamental for understanding the biomedical texts.

To fill the gap, many pre-trained language models for the biomedical domain such as BioBERT [15], BlueBERT [42], and ClinicalBERT [43] et al, have been proposed to further pre-train PLMs in the general domain with biomedical texts. BioBERT [15] is the first biomedical language model that further pre-trains BERT on biomedical scientific texts including PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC). Following it, BlueBERT [42] conducts continual pre-training on clinical text corpus MIMIC-III [44] along with biomedical scientific texts. Yet they still adopt the same vocabulary as BERT, which limits their ability in modeling the semantic information of biomedical texts. Different from them, SciBERT [45] builds a domain-specific vocabulary from scratch and conducts pre-training on scientific literature, in which 12% articles are from the computer science domain and 82% articles are from the biomedical domain. PubMedBERT [46] pre-trains their models with scientific papers solely in the biomedical domain. Compared with PLMs pre-trained in the general domain, biomedical PLMs



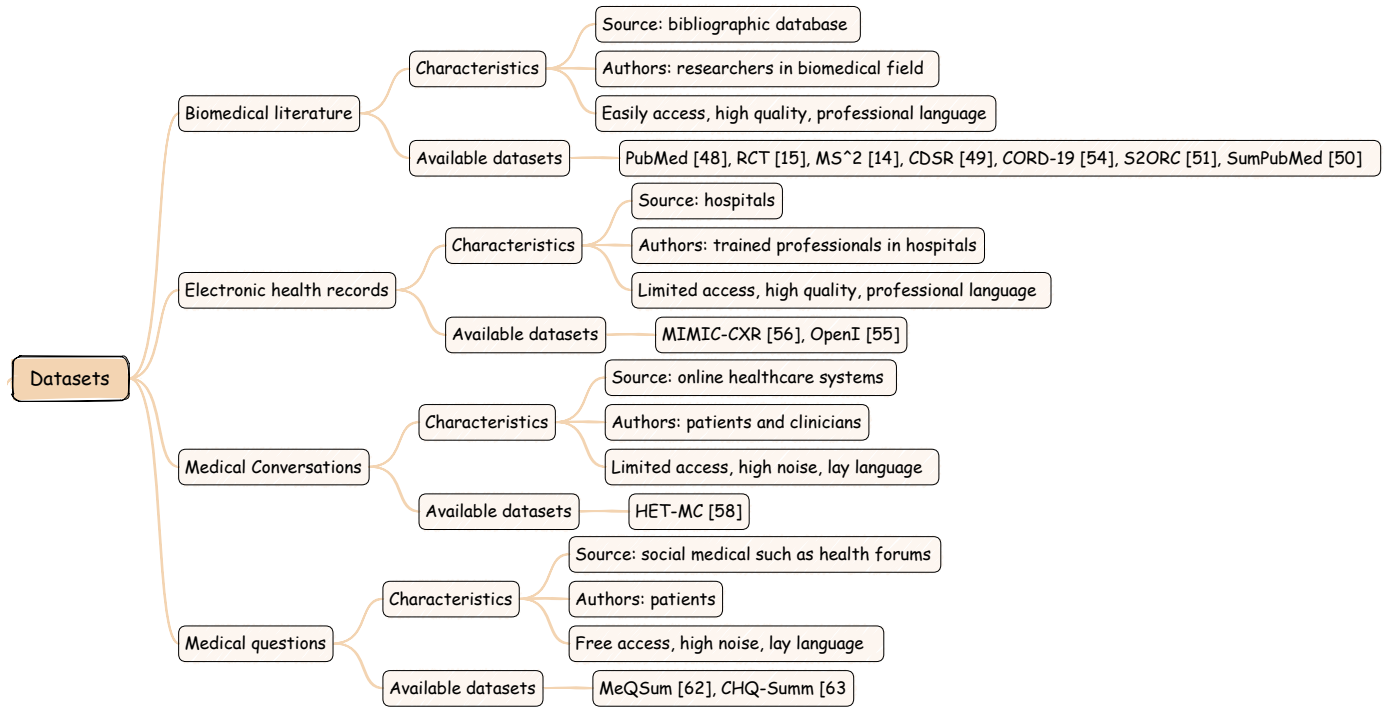


Fig. 5. Overview of datasets.

| Dataset             | Category              | Size    | Content                          | Summarization Task |
|---------------------|-----------------------|---------|----------------------------------|--------------------|
| PubMed [47]         | Biomedical literature | 133,215 | Full contents of articles        | Single             |
| SumPubMed [48]      | Biomedical literature | 33,772  | Full contents of articles        | Single             |
| S2ORC [49]          | Biomedical literature | 63,709  | Full contents of articles        | Single             |
| CORD-19 [50]        | Biomedical literature | -       | Full contents of articles        | Single             |
| CDSR [51]           | Biomedical literature | 7,805   | Abstracts of articles            | Single             |
| RCT [10]            | Biomedical literature | 4,528   | Titles and abstracts of articles | Multiple           |
| MS <sup>2</sup> [9] | Biomedical literature | 470,402 | Abstracts of articles            | Multiple           |
| MIMIC-CXR [52]      | EHRs                  | 124,577 | Full contents of reports         | Single             |
| OpenI [53]          | EHRs                  | 3,599   | Full contents of reports         | Single             |
| HET-MC [54]         | Medical conversation  | 109,850 | Multi-turn conversations         | Single             |
| MeQSum [55]         | Medical question      | 1,000   | Consumer health question         | Single             |
| CHQ-Summ [56]       | Medical question      | 1,507   | Full contents of question        | Single             |

TABLE 1  
Biomedical text summarization datasets.

based on biomedical texts can provide better representations that capture the contextual information of both normal tokens and domain-specific terms. One can check the survey paper [19] for more details on pre-trained language models for the biomedical domain.

### 3 DATASETS

Unstructured biomedical texts used in text summarization methods involve various types, including biomedical literature, electronic health records (EHRs), medical conversations, and medical questions, as shown in Figure 5. Details of these datasets are summarized in Table 1.

**Biomedical Literature** With the exponentially growing of scientific papers, developing automated summarization tools for biomedical articles has long attracted much attention. These texts are usually written by domain experts such as researchers and physicians. Compared with general domain texts such as social media texts or news texts,

they are less noisy and generally organized with standard sections, such as “Introduction”, “Methods”, “Results” et al.

For single document summarization, PubMed [47] is one of the most commonly used datasets, for summarization of long biomedical texts. It consists of 133K scientific papers collected from the PubMed open access repositories<sup>1</sup>. It has been used as a benchmark dataset and widely evaluated by the general text summarization methods and biomedical text summarization methods. It is noticed that Zhong et al [57] further adapt the dataset that only uses the introduction of texts as the input. To identify these two settings on the dataset, we name the original PubMed dataset that uses full contents of documents as the PubMed-Long<sup>2</sup>, and the new dataset that is adapted by Zhong et al [57] as the PubMed-Short. Following it, SumPubMed [48] proposed recently, includes 33,772 documents from Bio

1. <https://www.ncbi.nlm.nih.gov/pmc/tools/opaftlist/>  
 2. <https://github.com/armancohan/long-summarization>

Med Central (BMC) of PubMed archive<sup>3</sup>. Bishop et al [49], [58] extracts the subset from the large scientific corpus S2ORC [59] and build the dataset S2ORC<sup>4</sup> which includes 63,709 articles from the biological and biomedical domain. Most recently, COVID-19 Open Research Dataset (CORD-19<sup>5</sup>) [50] has attracted much attention, for which developing summarization systems would facilitate relevant research and help against the COVID-19 pandemic. CORD-19 contains millions of papers related to COVID-19, SARS-CoV-2, and other coronaviruses. Moreover, Guo et al [51] collected the CDSR<sup>6</sup> dataset to support the task of lay language summarization of biomedical scientific reviews, which is a special kind of single document summarization that aims to generate plain language abstract for lay people based on professional abstracts from expertise. It contains 7,805 abstract pairs of biomedical scientific reviews, in which professional abstracts of systematic reviews are deemed as inputs and their corresponding plain language abstracts as target summaries.

As for multi-document summarization in the biomedical domain, Wallace et al [10] build the RCT<sup>7</sup> summarization dataset with 4,528 data samples searched from PubMed<sup>8</sup>. The input of each data sample includes titles and abstracts of related papers describing randomized controlled trials (RCTs), while the conclusion section of the systematic review from Cochrane<sup>9</sup> is treated as the target summary. Similarly, Deyoung et al [9] developed the MS<sup>2</sup><sup>10</sup> for multi-document summarization of medical studies. It collected 470K papers from Semantic Scholar and 20K reviews that summarized these papers.

**Electronic Health Records** Electronic health records have been widely adopted by hospitals to store and manage medical information of patients, such as diagnostic codes, medications, laboratory results, clinical notes et al. They are also written by professionals with professional language and specific structure. Different from scientific papers that are generally free access, EHRs may have restrictions on public access due to privacy issues. Several publicly available datasets have been released to support the automated summarization of radiology reports. It is important to automatically generate the impression which should highlight key observations of the findings and background of the radiology reports. Demner et al [53] collected the OpenI<sup>11</sup> datasets containing 3,996 chest x-ray reports from hospitals within the Indiana Network. Compared with OpenI, MIMIC-CXR [52] is a larger publicly available dataset including 107,372 radiology reports from Beth Israel Deaconess Medical Center Emergency Department between 2011–2016<sup>12</sup>.

**Medical Conversations** Medical conversations between patients and doctors from online healthcare systems have

become an important source of medical information, with the increasing usage of telemedicine. The automated summarizing of key medical information on long medical conversations can save much time for doctors and improve healthcare efficiency. Medical conversations usually involve multi-turn interactions between two parties. The patients focus on asking questions and solutions to their health problems and describing their symptoms, while doctors would ask for detailed symptoms of patients and provide diagnostic suggestions. Similar to EHRs, accessing medical conversations at telemedicine platforms may have restrictions due to privacy concerns. Moreover, it is time-consuming and expensive to build the supervised data, since it requires professionals to write target summaries manually. Up to now, although several advanced methods with PLMs for medical conversation summarization have been proposed [54], [60], [61], [62], [63], publicly available datasets are limited. Song et al [54] proposed the Chinese medical conversation summarization dataset<sup>13</sup> with 109,850 conversations from the online health platform<sup>14</sup>.

**Medical Questions** The consumer health questions produced by healthcare consumers in the web such as health forums, are another important data source of clinical information. To find trustworthy answers for their health questions, healthcare consumers can query the web with long natural language questions with peripheral details. The peripheral information is useless to find high-quality answers for health questions. Therefore, summarizing consumer health questions into concise text with salient information is quite useful for improving efficiency of medical question answering. Abacha et al [55] build the MeQSum<sup>15</sup> corpus with 1,000 consumer health questions as inputs and their manual summaries from three medical experts. Yadav et al [56] introduced another dataset CHQ-Summ<sup>16</sup> most recently, which includes 1,507 consumer health questions and their summaries annotated by experts. Different from other texts such as biomedical papers with thousands of words, consumer health question-summary pairs are short texts. For example, in CHQ-Summ, the average length of questions and their summaries are 200 words and 15 words respectively.

## 4 PLMS FOR BIOMEDICAL TEXT SUMMARIZATION

Given biomedical text datasets, there are many methods that have been proposed to explore how to make better use of PLMs for the biomedical summarization task. Different from previous reviews [4], [16], [17], [18] for traditional and deep learning methods that classify methods according to their inputs (multiple document/single document) and outputs (extractive/abstractive), we focus on how PLMs are leveraged in recent research to improve biomedical text summarization. Thus, we first categorize them into three major categories: feature-based, fine-tuning-based, and domain-adaption-with-fine-tuning-based, according to the ways that they introduce PLMs into biomedical summarization. We

3. <https://github.com/vgupta123/sumpubmed>

4. <https://github.com/jbshp/GenCompareSum>

5. <https://github.com/allenai/cord19>

6. [https://github.com/qiuweipku/Plain\\_language\\_summarization](https://github.com/qiuweipku/Plain_language_summarization)

7. <https://github.com/bwallace/RCT-summarization-data>

8. <https://pubmed.ncbi.nlm.nih.gov>

9. <https://www.cochranelibrary.com/>

10. <https://github.com/allenai/ms2/>

11. <https://openi.nlm.nih.gov/faq#collection>

12. <https://physionet.org/content/mimic-cxr/2.0.0/>

13. <https://github.com/cuhksz-nlp/HET-MC>

14. <https://www.chunyuyisheng.com/>

15. <https://github.com/abachaa/MeQSum>

16. <https://github.com/shwetantp/Yahoo-CHQ-Summ>

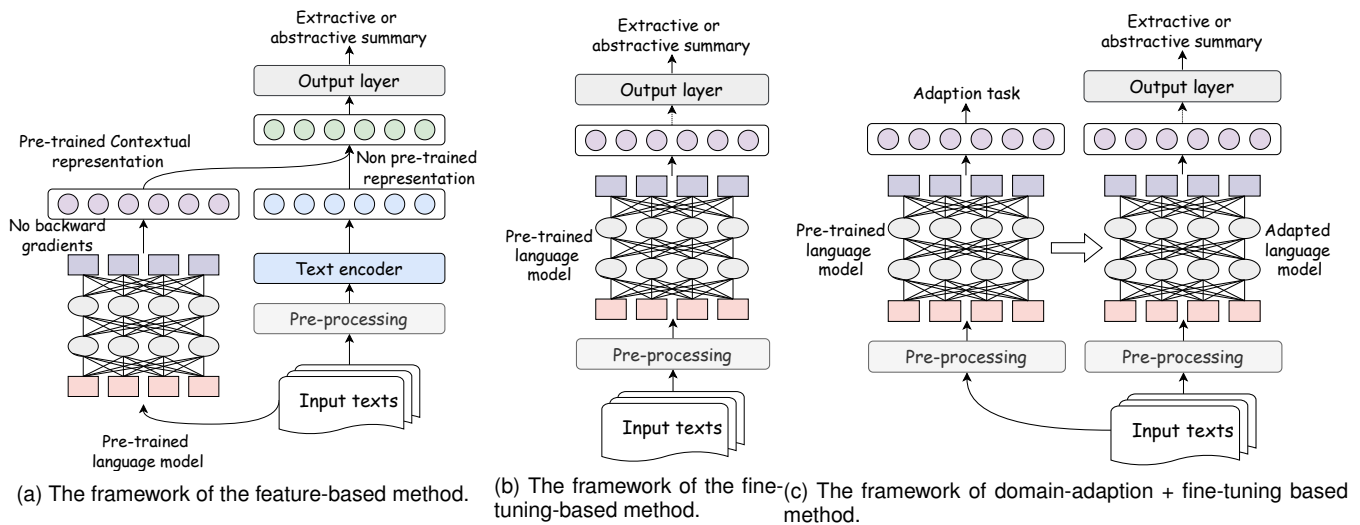


Fig. 6. Comparison of different strategies on using pre-trained language models.

then further categorize methods according to the structure of PLMs and the dataset they adopt.

As shown in Figure 6, the feature-based methods independently utilize contextual representations from PLMs without refining the pre-trained parameters of PLMs. The representations from PLMs are generally concatenated with the representations from the encoder to generate the output. Although they are time-saving for fixing the parameters of PLMs, the performance is limited without considering the task-specific supervised information. The fine-tuning-based methods generally take PLMs as the text encoder whose generated representations are directly fed into the output layer. The parameters of PLMs along with task-specific parameters are fine-tuned according to the task loss. Compared with feature-based methods, they usually require much higher computing resources and are time-consuming but can yield improved and promising performance. The fine-tuning-with-domain-adaption-based methods first conduct the domain-adaption for PLMs via continually pre-training PLMs with designed tasks on the target data and then fine-tune the adapted PLMs along with task-specific layers. The domain adaption allows PLMs to better capture domain-specific knowledge compared with PLMs only pre-trained on the general domain texts, resulting in a better performance on the summarization task.

Next, we will review and discuss these methods in more detail, as shown in Table 2.

#### 4.1 Feature-based Methods

Feature-based methods take contextualized representations of PLMs as the extra features, which can provide contextual semantic information that is ignored by the traditional text encoder.

**Bi-directional encoder language models** For biomedical literature texts, Moradi et al [64] propose the unsupervised extractive summarizer based on hierarchical clustering and PLMs. They conduct sentence clustering based on sentence representations from PLMs and then select top sentences from clusters with the ranking method. They investigate

different versions of BERT and BioBERT to yield sentence representations. The proposed methods show better performance than traditional unsupervised methods such as TextLexAn<sup>17</sup>. They find that all versions of BioBERT (BioBERT-pmc, BioBERT-pubmed, BioBERT-pubmed+pmc) outperform the BERT-base, but underperform the BERT-large. Moradi et al [65] propose the graph ranking based method for biomedical text summarization. They use the contextualized embeddings of BioBERT to represent sentences and build graphs for texts. The important sentences are identified with the graph ranking algorithm from text graphs. The model based on BioBERT-pubmed+pmc achieves better performance than models based on the other two versions of BioBERT: BioBERT-pmc and BioBERT-pubmed.

As for clinical notes, Gharebagh et al [66] develop supervised method for abstractive summarization of clinical notes. They propose to incorporate the contextual embeddings from BERT as the input embeddings of the Bi-LSTM-based encoder. It shows better performance than classical sequence-to-sequence-based methods based on recurrent neural networks. Yan et al [67] propose the radiology-specialized language model RadBERT that is pre-trained on millions of radiology reports. On the unsupervised extractive summarization of radiology reports, it achieves better performance than other language models including BERT, BioBERT, ClinicalBERT, BlueBERT, and BioMed-RoBERTa [88]. Among all variants: RadBERT-BERT-base, RadBERT-RoBERTa, RadBERT-ClinicalBERT et al, the RadBERT-BioMed-RoBERTa achieves the best performance.

**Auto-regressive decoder language models** Bishop et al [49] present unsupervised extractive summarization method GenCompareSum for biomedical literature. GenCompareSum uses the T5 generative model to generate key snippets for text sections and selects important sentences with BERTScore [89] between key snippets and sentences. It outperforms traditional unsupervised methods such as LexRank [90], TextRank [11], and also the SOTA supervised

17. <http://texlexan.sourceforge.net>



| Paper                   | Strategy                  | Model                    | Category       | Input    | Output      | Training      | Data                     |
|-------------------------|---------------------------|--------------------------|----------------|----------|-------------|---------------|--------------------------|
| Moradi et al [64]       | feature-base              | encoder                  | literature     | single   | extractive  | unsupervised  | -                        |
| Moradi et al [65]       | feature-base              | encoder                  | literature     | single   | extractive  | unsupervised  | -                        |
| Gharebagh et al [66]    | feature-base              | encoder                  | EHRs           | single   | abstractive | supervised    | MIMIC-CXR, OpenI         |
| RadBERT [67]            | feature-base              | encoder                  | EHRs           | single   | extractive  | unsupervised  | -                        |
| GenCompareSum [49]      | feature-base              | decoder                  | literature     | single   | extractive  | unsupervised  | PubMed, CORD-19, S2ORC   |
| Su et al [68]           | fine-tuning, feature-base | encoder, encoder-decoder | literature     | multiple | hybrid      | un-supervised | CORD-19                  |
| ContinualBERT [69]      | fine-tuning               | encoder                  | literature     | single   | extractive  | supervised    | CORD-19                  |
| BioBERTSum [70]         | fine-tuning               | encoder                  | literature     | single   | extractive  | supervised    | -                        |
| Cai et al [71]          | fine-tuning               | encoder                  | literature     | single   | abstractive | supervised    | CORD-19, PubMed          |
| Kanwal et al [72]       | fine-tuning               | encoder                  | EHRs           | single   | extractive  | unsupervised  | MIMIC-III                |
| Hu et al [73]           | fine-tuning               | encoder                  | EHRs           | single   | abstractive | supervised    | MIMIC-CXR, OpenI         |
| HET [54]                | fine-tuning               | encoder                  | conversation   | single   | extractive  | supervised    | HET-MC                   |
| Esteve et al [6]        | fine-tuning               | decoder                  | literature     | multiple | abstractive | supervised    | CORD-19                  |
| Chintagunta et al [61]  | fine-tuning               | decoder                  | conversation   | single   | abstractive | supervised    | -                        |
| Deyoung et al [9]       | fine-tuning               | encoder-decoder          | literature     | multiple | abstractive | supervised    | MS <sup>2</sup>          |
| Zhu et al [74]          | fine-tuning               | encoder-decoder          | EHRs           | single   | abstractive | supervised    | MIMIC-CXR, OpenI         |
| Kondadadi et al [75]    | fine-tuning               | encoder-decoder          | EHRs           | single   | abstractive | supervised    | MIMIC-CXR, OpenI         |
| Xu et al [76]           | fine-tuning               | encoder-decoder          | EHRs, question | single   | abstractive | supervised    | MIMIC-CXR, OpenI, MeQSum |
| He et al [77]           | fine-tuning               | encoder-decoder          | EHRs, question | single   | abstractive | supervised    | MIMIC-CXR, OpenI, MeQSum |
| Yadav et al [78]        | fine-tuning               | encoder-decoder          | question       | single   | abstractive | supervised    | MeQSum                   |
| CLUSTER2SENT [79]       | fine-tuning               | encoder-decoder          | conversation   | single   | abstractive | supervised    | -                        |
| Zhang et al [62]        | fine-tuning               | encoder-decoder          | conversation   | single   | abstractive | supervised    | -                        |
| Navarro et al [80]      | fine-tuning               | encoder-decoder          | conversation   | single   | abstractive | supervised    | -                        |
| BioBART [81]            | fine-tuning               | encoder-decoder          | conversation   | single   | abstractive | supervised    | -                        |
| KeBioSum [58]           | adaption+fine-tuning      | encoder                  | literature     | single   | extractive  | supervised    | PubMed, CORD-19, S2ORC   |
| Yalunin et al [82]      | adaption+fine-tuning      | encoder                  | EHRs           | single   | abstractive | supervised    | -                        |
| Mahajan et al [83]      | adaption+fine-tuning      | encoder                  | EHRs           | single   | abstractive | supervised    | MIMIC-CXR, OpenI         |
| Yadav et al [84]        | adaption+fine-tuning      | encoder                  | question       | single   | abstractive | supervised    | MeQSum                   |
| Kieuvongngam et al [85] | adaption+fine-tuning      | decoder                  | literature     | single   | hybrid      | supervised    | CORD-19                  |
| Guo et al [51]          | adaption+fine-tuning      | encoder, encoder-decoder | literature     | single   | hybrid      | supervised    | CDSR                     |
| Wallace et al [10]      | adaption+fine-tuning      | encoder-decoder          | literature     | multiple | abstractive | supervised    | RCT                      |
| BDKG [86]               | adaption+fine-tuning      | encoder-decoder          | EHRs           | single   | abstractive | supervised    | MIMIC-CXR, OpenI         |
| Mrini et al [87]        | adaption+fine-tuning      | encoder-decoder          | question       | single   | abstractive | supervised    | MeQSum                   |

TABLE 2

Overview of Methods. Except how they use PLMs and what PLMs they used, we also list other features of these methods, according to types of output summary (extractive, abstractive, hybrid), numbers of input documents (single, multiple), and types of input documents (biomedical literature summarization, radiology report summarization, medical dialogue summarization, and medical question summarization). "-" means datasets that are not released.

method BERTSum.

**Encoder-decoder language models** Su et al [68] propose the query-focused multi-document summarizer for COVID-19 articles, that is able to generate abstractive and extractive summaries based on user queries. For extractive summarization, they proposed the feature based method with ALBERT [91], to generate sentence representations and calculate the cosine similarity between sentences and queries to select important sentences.

## 4.2 Fine-tuning based Methods

Fine-tuning is the most common way to use PLMs for downstream tasks. Different from feature-based methods that freeze the parameters of PLMs, fine-tuning-based methods refine all parameters of PLMs along with task-specific parameters.

**Bi-directional encoder language models** For biomedical literature, Park et al [69] present the ContinualBERT model for adaptive extractive summarization of covid-19 related literature. ContinualBERT trains two BERT models with continual learning in order to process texts online. It shows better performance than the SOTA extractive method BERTSum [92]. Du et al [70] propose BioBERTSum model for extractive summarization of biomedical literature. It uses token embedding, sentence embedding, and position embedding to embed input texts, and then yields contextual representations of sentences with BioBERT. BioBERT and the

extra classifier layer are fine-tuned with the cross-entropy loss. It proves the advantage of using a domain-specific language model for biomedical texts and outperforms the SOTA method BERTSum that uses BERT in the general domain as the encoder. Cai et al [71] propose a SciBERT-based abstractive summarization model for COVID-19 scientific papers, that uses the linguistic information of word co-occurrence encoded by graph attention network to enrich the SciBERT encoder. They find that their method based on the SciBERT encoder outperforms that based on BERT and BioBERT encoder.

As for clinical notes, Kanwal et al [72] propose the multi-head attention-based method for extractive summarization. It fine-tunes BERT on the task of predicting and identifying ICD-9 labels on the ICD-9 labeled MIMIC-III discharge notes. The attention scores of sentences from the last layer of the BERT model are used to select sentences. Hu et al [73] propose the radiology report summarizer that uses BioBERT as text encoder and randomly initialized transformer layers as the decoder. They use the graph encoder and contrastive learning to incorporate extra knowledge to improve the BioBERT encoder. The proposed method achieves the SOTA in the radiology report summarization.

Moreover, Song et al [54] propose the hierarchical encoder-tagger (HET) model for extractive summarization of medical conversation, which includes token-level and utterance-level encoders to encode input long transcripts. They use the Chinese version of BERT as the token-level

encoder.

**Auto-regressive decoder language models** There are abstractive methods that focus on fine-tuning language models such as GPT-2 and GPT-3. They are generally based on the classical encoder-decoder framework and take auto-regressive decoder language models as the text decoder. For COVID-19 related biomedical articles, Esteva et al [6] design the parallel encoder-decoder framework for abstractive summarization of multiple COVID-19 articles, that uses BERT as encoder and GPT-2 as the decoder. To overcome the problem of limited training data, methods are proposed to use the few-shot learner such as GPT-3 or few-shot fine-tuning strategy.

The auto-regressive decoder language models are also introduced for medical dialogue summarization. Chintagunta et al [61] integrate medical knowledge and GPT-3. They consider GPT-3 as the summary generator and choose the best summary that captures the most medical concepts. They show that GPT-3 can be a promising backbone method for generating high-quality training data that can be incorporated with the training data with human annotation.

**Encoder-decoder language models** Encoder-decoder language models such as BART, T5, and PEGASUS [93] that are pre-trained with an objective function specifically for abstractive text summarization, have been widely used for biomedical abstractive summarization.

For biomedical literature, Deyoung et al [9] develop the BART-based method for multi-document summarization of medical studies. To encode multi-documents, they investigate two encoders. One is using multiple BART encoders to encode multi-documents separately. Another one is using LongformerEncoderDecoder (LED) [94], which can encode long inputs up to 16K tokens. Su et al [68] also propose the query-focused multi-document summarizer for COVID-19 articles, that is able to generate abstractive and extractive summaries based on user queries. They fine-tune BART for multi-document abstractive summarization.

As for EHRs, the MEDIQA 2021 Shared Task [95] at the BioNLP 2021 workshop introduces the abstractive summarization task for radiology reports and medical question, in which most participating teams propose methods based on encoder-decoder language models. Among 14 teams that participate in the radiology reports summarization task, 6 of them [74], [75], [76], [77], [83] use the encoder-decoder language models such as BERT, BART, PEGASUS [93]. Most of them [74], [76] find that fine-tuning PEGASUS achieves the best performance, while Kondadadi et al [75] reports that the best performance is achieved by the BART. Moreover, they report that adapting PEGASUS on the PubMed corpus can lead to worse performance, which may be due to the gap between biomedical literature and medical reports.

For medical conversation summarization, efforts have been proposed to explore the encoder-decoder language models to address challenges such as limited labeled data and long transcripts. Krishna et al [79] develop CLUSTER2SENT, an extractive-abstractive hybrid method on doctor-patient conversations to generate SOAP notes (long semi-structured clinical summaries). T5 model is used in the abstractive module of CLUSTER2SENT. Zhang et al [62] leverage the BART model for automatic summarization of doctor-patient conversations with limited labeled training

data. They propose the multistage fine-tuning strategy to address the input length limitation of BART. They find that fine-tuning BART can generate summaries of good quality even with limited training data. Moreover, they also find that the BART-based model significantly outperforms the BigBird [96] based models that are initialized by RoBERTa-base and PEGASUS-Large. Navarro et al [80] explore fine-tuning BART, T5, PEGASUS with zero-shot and few-shot learning strategies for medical dialogue summarization with small training data. They find that BART achieves the best performance among these PLMs. Yuan et al [81] develop the first encoder-decoder pre-trained language model BioBART in the biomedical domain, which has shown better performance on medical dialogue summarization than BART.

Moreover, for medical questions, 22 teams participate in the medical question summarization task of the MEDIQA 2021 Shared Task [95] at the BioNLP 2021 workshop, and all methods are based on fine-tuning encoder-decoder language models. The best performance is achieved by the ensemble model [77] that re-ranks summary outputs of multiple advanced encoder-decoder language models including BART, T5, and PEGASUS. Yadav et al [78] present the reinforcement learning based framework for abstractive summarization of medical questions. They propose two reward functions: the Question-type Identification Reward (QTR) and Question-focus Recognition Reward (QFR), which are optimized via learning optimal policy defined by BERT. They show that the encoder-decoder language model ProphetNet [97] with the proposed reward functions has better performance than other PLMs including T5, BART, and PEGASUS.

### 4.3 Domain Adaption with Fine-tuning based Methods

There are summarization methods that conduct domain adaption before fine-tuning PLMs, to capture domain and task-specific information.

**Bi-directional encoder language models** For biomedical literature, Xie et al [58] propose the KeBioSum for the extractive summarization. It proposes to refine PLMs with the domain adaption tasks of predicting key entities and their types based on the lightweight fine-tuning framework, which aims to incorporate fine-grained medical knowledge into PLMs. It proves that although biomedical language models such as BioBERT, and PubMedBERT can capture domain knowledge to some extent, fine-grained medical knowledge is still beneficial to improve language models. They find that PubMedBERT-based methods outperform methods based on BERT, RoBERTa, and BioBERT.

As for EHRs, Yalunin et al [82] present the abstractive summarizer for patient hospitalization histories, that uses Longformer [94] as the encoder and BERT as the decoder. They propose to pre-train BERT and Longformer with the masked language task on the hospitalization history dataset before task specific fine-tuning. Dai et al [86] proposed the BDKG method, which achieves the best performance on MEDIQA 2021 Shared Task for radiology reports summarization. It ensembles results from multiple language models BART, DistillBERT [98], PEGASUS, and uses other strategies including domain adaption and text normalization.

| Dataset        | Category     | Metrics  |                                       |                  |                           |                                |   |
|----------------|--------------|--|---------------------------------------|------------------|---------------------------|--------------------------------|---|
|                |              | Automatic metrics  |                                       |                  | Human evaluations         |                                |   |
|                |              | similarity   | factuality                            | others           | similarity                | factuality                     | others                                      |
| PubMed [47]    | literature   | ROUGE [49], [58], [69], [71], BERTScore [58]   | -                                     | -                | -                         | -                              | -   |
| SumPubMed [48] | literature   | ROUGE [48]   | -                                     | -                | -                         | -                              | -   |
| S2ORC [49]     | literature   | ROUGE [49], [58], [69], [71], BERTScore [58]   | -                                     | -                | -                         | -                              | -   |
| CORD-19 [50]   | literature   | ROUGE [49], [58], [69], [71], BERTScore [58]   | -                                     | -                | -                         | -                              | -   |
| CDSR [51]      | literature   | ROUGE [51]   | -                                     | readability [51] | meaning preservation [51] | correctness [51]               | understandability [51], grammaticality [51] |
| RCT [10]       | literature   | ROUGE [10]   | -                                     | -                | relevance [10]            | factuality [10]                | fluency [10]                                |
| MS-2 [9]       | literature   | ROUGE [9]  | $\Delta EI$ [9]                       | -                | -                         | PICO, direction, modality [99] | grammar, lexical, non-redundancy [99]       |
| MIMIC-CXR [52] | EHRs         | ROUGE [66], [73], [74], [75], [76], [77], [83], [86], BERTScore [76], [77], [83], [86] | CheXbert [73], [76], [77], [83], [86] | -                | -                         | -                              | -   |
| OpenI [53]     | EHRs         | ROUGE [66], [73], [74], [75], [76], [77], [83], [86], BERTScore [76], [77], [83], [86] | CheXbert [73], [76], [77], [83], [86] | -                | -                         | -                              | -   |
| HET-MC [54]    | conversation | ROUGE [54]   | -                                     | -                | -                         | -                              | -   |
| MeQsum [55]    | question     | ROUGE [78], [84], [87]   | -                                     | -                | semantics preserved [78]  | factual consistent [78]        | -   |
| CHQ-Summ [56]  | question     | ROUGE, BERTScore [56]  | -                                     | -                | -                         | -                              | -   |

TABLE 3  
The usage of evaluation metrics on different biomedical datasets and methods.

Moreover, for medical questions, Yadav et al [84] investigate to incorporate the knowledge of "question-focus" and "question-type" with PLMs for abstractive summarization of consumer health questions. To induce PLMs to capture this knowledge, they adapt PLMs with designed Cloze tasks.

**Auto-regressive decoder language models** Kieu-vongngam et al [85] use the GPT-2 [35] for abstractive summarization of COVID-19 medical research articles. They take keywords of articles as inputs and fine-tune GPT-2 on multi-tasks including the language modeling task and the multiple choice prediction task.

**Encoder-decoder language models** For biomedical articles, Guo et al [51] use the BART for automated lay language summarization of biomedical review articles. They conduct domain-adaption before fine-tuning, which pre-trains the BART model to reconstruct original PubMed abstracts with disrupted abstracts. Wallace et al [10] propose the multi-document abstractive summarization models based on BART for randomized controlled trials (RCTs). They adapt the BART with the domain-specific pre-training strategy of generating summaries from full-text articles before fine-tuning. They also use the "decoration" strategy to explicitly inform key trial attributes (the "PICO" elements) of input articles.

As for medical questions, Mrini et al [87] present the multi-task learning and data augmentation method on medical question summarization and recognizing question entailment (RQE) for medical question understanding. They prove that the multi-task learning between question summarization and RQE is able to increase the performance of PLMs including BART and BioBERT.

## 5 EVALUATIONS

Evaluating summaries of biomedical texts is more challenging than general document summarization since biomedical texts are more technical and complex than general texts in length and structure. As shown in Table 3, we first categorize metrics into automatic metrics and human evaluation according to whether human efforts are involved. Automatic metrics for biomedical text summarization can verify the performance of methods without any human effort. However, it can only leverage the shallow lexical and syntactical information of the generated and gold summaries. In contrast, human evaluation can capture and model semantic information which is generally difficult to quantize. Nevertheless, human evaluation in the biomedical

domain which requires specific domain knowledge has an even higher time and financial consumption than it in the general domain.

We further divide existing metrics into three different classes, consisting of similarity, factuality, and others. Similarity metrics focus on the relevance of generated summaries with gold summaries, which are generally based on the overlapping of tokens, phrases, and sentences. Factuality metrics verify the factual agreement of generated summaries with original documents, which is a critical measurement for the real application of automatic systems, especially in the biomedical domain. Moreover, there are other metrics such as: 1) interpretability that tests whether human readers can easily understand generated summaries, 2) fluency: how fluent and coherent generated summaries are, and 3) grammaticality: how grammatically correct generated summaries are.

### 5.1 Automatic metrics

**Similarity** Similar to the general domain, ROUGE [100] is the most widely used metric for biomedical summarizers, including (1) ROUGE-1: unigram overlap between generated summaries of summarizers and gold summaries; (2) ROUGE-2: bigram overlap between generated summaries of summarizers and gold summaries; and (3) ROUGE-L: the longest common subsequences between generated summaries of summarizers and gold summaries. All biomedical text summarization datasets adopt the ROUGE score to evaluate their performance. However, ROUGE metrics are limited to relying on shallow lexical overlaps without considering the paraphrasing and terminology variations when measuring similarity. To address this issue, there is a recently proposed BERTScore [89] metric that is also introduced to several common datasets. It calculates the similarity between two sentences as the sum of cosine similarities between the contextual embeddings of their tokens from pre-trained language models.

**Factuality** Compared with extractive methods, it is reported [101] that abstractive methods struggle to generate factual correct summaries. It has been a growing awareness that metrics such as ROUGE and BERTScore can not reflect the factual correctness of generated summaries [102]. For automatic evaluation, Deyoung et al [9] propose the  $\Delta EI$  metric to calculate the factual agreement of generated summaries and input medical studies. They propose to calculate the Jensen-Shannon Distance (JSD) between distributions of generated summaries and input medical studies in three

directions (increase, decrease, no change) of reported directionality. However, the metric is only adopted in the MS<sup>2</sup> dataset and there are no efforts in other biomedical literature datasets considering the factual correctness. Zhang et al [102] propose the factual F1 score to evaluate the factual correctness of generated summaries of radiology reports. They propose to use the CheXbert labeler [103] to yield the binary presence values of disease variables of generated summaries and references and then calculate the overlap of yielded binary presence values between them.

**Others** Guo et al [51] propose to apply the readability evaluation which verifies if the generated summaries are understandable for laymen. It utilizes three different standards including Flesch-Kincaid grade level [104], Gunning fog index [105], and Coleman-Liau index [106]. The readability metric is specially designed for their summarization task and dataset CDSR.

## 5.2 Human evaluation

Although manually evaluating the performance of summarization methods is time-consuming and expensive, human evaluation with domain experts can capture more aspects than automatic evaluation, such as fluency, coherence, factuality, and grammaticality et al. Generally, the human evaluation would recruit human evaluators which are able to read and write in English and participate in medical training and biology courses. Evaluators are required to score the generated summaries with designed questions that focus on one of the aforementioned aspects. Moramarco et al [107] study the correlation between human evaluation and 18 automatic evaluation metrics including text overlap metrics such as ROUGE, CHRF [108], embedding metrics BertScore, and factual F1 score et al, on generated clinical consultation notes. They find that simple character-based metrics such as character-based Levenshtein distance can be more effective than other complex metrics such as BERTScore, and the choice of human references can largely influence the performance of automatic evaluation metrics.

**Similarity** For biomedical literature, Guo et al [51] also propose to use human evaluation to further assess the quality of generated summaries on the CDSR dataset. Different from other research, it requires human evaluators to not participate in medical training and biology courses since its method is designed for laymen. It proposes the meaning preservation metric for human evaluators, which requires them to answer whether the generated summaries cover the key information of source documents, on the 1-5 Likert scale (1 means very poor, 5 means very good). For multi-document summarization dataset RCT, Wallace et al [10] ask evaluators who are medical doctors to score the relevance of the generated summaries to the given topic from mostly off-topic to strongly on-topic.

As for medical questions, Yadav et al [78] request two experts in medical informatics to measure the semantics preserved in the generated summaries, i.e, whether the question intent was mentioned in the generated summary on both MEQSUM and MATINF.

**Factual consistency** For biomedical literature, Guo et al [51] propose a correctness question for key information which requires evaluators to judge the quality of the generated summaries on the CDSR dataset following the same

1-5 Likert scale. Wallace et al [10] also request evaluators to answer two questions about the factuality of the generated summaries for RCT which concern the directory and the degree of the generated summaries compared with gold summaries. Otmakhova et al [99] define three different metrics for the factuality of the generated summaries on the MS<sup>2</sup> dataset, including (1) PICO correctness: the generated summary should contain the same patient population, intervention, and outcome (which are the entity types defined by PICO) as the gold summary; (2) direction correctness: the generated summary should have the same direction referring to the intervention's effect to the outcome as the gold summary, which can be classified as positive effect, negative effect, and no effect; (3) modality: the confidence of the generated summary about the claim should be the same as the gold summary, which can be defined as strong claim, moderate claim, weak claim, no evidence, and no claim.

As for medical questions, Yadav et al [78] also ask experts to verify if all key entities appear in the generated summaries on RCT as factual consistent.

**Other** There are also several methods that propose new human evaluation metrics to help understand the performance of biomedical summarizers. Moramarco et al [109] propose an objective human evaluation based on counting medical facts for generated summaries of medical reports. Guo et al [51] design two questions for grammatical correctness and readability respectively for CDSR dataset. Wallace et al [10] ask evaluators to evaluate the fluency of the generated summaries on the RCT dataset. Otmakhova et al [99] consider the fluency of the generated summaries and propose three metrics to evaluate it, including grammatical correctness, lexical correctness, and absence of repetition.

## 6 DISCUSSION

In this section, we make a further discussion on existing methods and their limitations, and then outlook promising future directions.

### 6.1 Comparison

We first present the performance of existing SOTA PLMs based methods on different datasets in Table 4, Table 5, Table 6, Table 7, Table 8, and Table 9.

**How do PLMs in biomedical text summarization work?** For biomedical literature, on the PubMed dataset, as shown in Table 4 and Table 5, PLMs-based methods such as PubMedBERTSum [92], and KeBioSum [58] outperforms TextRank [11] without PLMs on PubMed-long, and GenCompareSum [49], BERTSum [92], MatchSum [57], and KeBioSum [58] show a significant improvement in both ROUGE metrics compared with TextRank without PLMs on PubMed-short. It is noticed that the PubMed dataset is also used as a benchmark dataset in the general domain. Although there are other advanced methods such as LongT5 [111] have achieved the new SOTA on the dataset, we mainly focus on comparing methods that are designed for the biomedical domain in here. As for CORD-19 and S2ORC, Table 6 shows that GenCompareSum [49] and BERTSum [92] present great performance compared with existing non PLMs-based methods such as TextRank

| Methods            | Strategy                    | Model      | ROUGE-1 | ROUGE-2 | ROUGE-L |
|--------------------|-----------------------------|------------|---------|---------|---------|
| TextRank [11]      | -                           | -          | 34.53   | 12.98   | 30.99   |
| BERTSum [92]       | fine-tuning                 | BERT       | 34.00   | 13.42   | 30.69   |
| PubMedBERTSum [92] | fine-tuning                 | PubMedBERT | 34.98   | 14.22   | 31.37   |
| KeBioSum [58]      | domain adaption+fine-tuning | PubMedBERT | 36.39   | 16.27   | 33.28   |

TABLE 4

ROUGE F1 score of generated summaries by the SOTA extractive methods on the PubMed-long dataset, that extract 3 sentences to formulate the final summary.

| Methods       | Strategy                    | Model      | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---------------|-----------------------------|------------|---------|---------|---------|
| TextRank [11] | -                           | -          | 38.15   | 12.99   | 34.77   |
| BERTSum [92]  | fine-tuning                 | BERT       | 41.09   | 15.51   | 36.85   |
| MatchSum [57] | fine-tuning                 | RoBERTa    | 41.21   | 14.91   | 36.75   |
| KeBioSum [58] | domain adaption+fine-tuning | PubMedBERT | 43.98   | 18.27   | 39.93   |

TABLE 5

ROUGE F1 score of generated summaries by the SOTA extractive methods on the PubMed-short dataset, that extract 6 sentences to formulate the final summary.

| Data               |              |       | CORD-19 |         |         | S2ORC   |         |         |
|--------------------|--------------|-------|---------|---------|---------|---------|---------|---------|
| Methods            | Strategy     | Model | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-1 | ROUGE-2 | ROUGE-L |
| TextRank [11]      | -            | -     | 32.99   | 10.39   | 24.471  | 36.58   | 13.23   | 33.10   |
| SumBasic [110]     | -            | -     | 33.88   | 8.24    | 30.86   | 36.63   | 10.43   | 33.68   |
| BERTSum [92]       | fine-tuning  | BERT  | 38.95   | 12.17   | 35.48   | 43.56   | 17.85   | 40.40   |
| GenCompareSum [49] | feature-base | T5    | 41.02   | 13.79   | 37.25   | 43.39   | 16.84   | 39.82   |

TABLE 6

ROUGE F1 score of generated summaries by the SOTA extractive methods on the CORD-19 and S2ORC datasets, which extract 8 and 9 sentences for the CORD-19 and S2ORC datasets correspondingly to formulate the final summary.

| Methods  | Strategy    | Model      | R-1   | R-2  | R-L   | PICO | Direction | Modality | Grammar | Lexical | Non-redundancy |
|----------|-------------|------------|-------|------|-------|------|-----------|----------|---------|---------|----------------|
| BART [9] | fine-tuning | BART       | 27.56 | 9.40 | 20.80 | 45%  | 77%       | 45%      | 75%     | 69%     | 85%            |
| LED [9]  | fine-tuning | Longformer | 26.89 | 8.91 | 20.32 | 40%  | 75%       | 44%      | 63%     | 73%     | 89%            |

TABLE 7

ROUGE F1 score [100], factual correctness, grammatical errors and fluency [99] of generated summaries by SOTA abstractive methods for multiple biomedical document summarization on MS<sup>2</sup>. ROUGE-1, ROUGE-2 and ROUGE-L are commonly used for evaluating the relevancy between gold summaries and generated summaries. PICO, direction, and modality are used for evaluating the factual correctness of generated summaries.

| Metrics             |             |         | MIMIC-CXR |         |         | OpenI   |         |         |
|---------------------|-------------|---------|-----------|---------|---------|---------|---------|---------|
| Method              | Strategy    | Model   | ROUGE-1   | ROUGE-2 | ROUGE-L | ROUGE-1 | ROUGE-2 | ROUGE-L |
| LexRank [90]        | -           | -       | 18.11     | 7.47    | 16.87   | 14.63   | 4.42    | 14.06   |
| TransformerEXT [92] | -           | -       | 31.00     | 16.55   | 27.49   | 15.58   | 5.28    | 14.42   |
| OntologyABS [66]    | -           | -       | 53.57     | 40.78   | 51.81   | -       | -       | -       |
| Hu et al [73]       | fine-tuning | BioBERT | 57.38     | 45.52   | 56.13   | 54.52   | 64.97   | 55.59   |

TABLE 8

ROUGE F1 score of generated summaries by the SOTA abstractive methods for radiology findings summarization on MIMIC-CXR and OpenI.

| Method           | Strategy                    | Model   | ROUGE-1 | ROUGE-2 | ROUGE-L |
|------------------|-----------------------------|---------|---------|---------|---------|
| Seq2Seq [24]     | -                           | -       | 25.28   | 14.39   | 24.64   |
| BertSum [92]     | fine-tuning                 | BERT    | 26.24   | 16.20   | 30.59   |
| PEGASUS [93]     | fine-tuning                 | PEGASUS | 39.06   | 20.18   | 42.05   |
| Yadav et al [84] | domain adaption+fine-tuning | MiniLM  | 45.20   | 28.38   | 48.76   |
| Mrini et al [87] | domain adaption+fine-tuning | BART    | 54.5    | 37.9    | 50.2    |

TABLE 9

ROUGE F1 score of generated summaries by the SOTA abstractive methods on the MeQSum dataset.



and SumBasic [110]. Moreover, for multiple document summarization dataset MS<sup>2</sup>, as shown in Table 7, PLMs-based methods such as BART [9] and LED [9] have shown great performance in both automatic metrics and human evaluation. We can see that they can generate summaries that have low grammar and lexical error (less than 31%), and low redundancy (less than 15%). However, they have problem on generating factual correct summaries, for example with PICO correctness no higher than 45%.

For EHRs, Table 8 clearly proves the benefits of PLMs via the comparison of PLMs-based methods such as Hu et al [73] and traditional methods such as OntologyABS [66], LexRank, and TransformerEXT [92]. Hu et al [73] present the best performance on both MIMIC-CXR and OpenI datasets. Moreover, as shown in Table 9, PLMs-based methods such as BertSum, PEGASUS [93], Yadav et al [84], and Mrini et al [87] both outperform traditional Seq2Seq [24] method on the medical question dataset MeQSum.

Overall, we can find that performance on various datasets has been greatly boosted by advanced methods that make use of PLMs, and existing methods are able to generate fluent summaries. However, their performances are still far away from desirable especially in factuality.

**What is the optimal way to introduce PLMs in biomedical text summarization?** From Table 5 and Table 9, we can find that domain-adaption with fine-tuning based methods such as KeBioSum [112] and Mrini et al [87] present the best performance among three different ways to introduce PLMs. We believe it should also have a superior performance with domain-adaption with fine-tuning mechanism to further capture domain knowledge, although there are no previous efforts to apply it to the multiple document summarization dataset MS<sup>2</sup> and EHRs datasets.

Fine-tuning based methods generally yield better performance than feature-based methods, except for methods on the CORD-19 and S2ORC datasets, as shown in Table 6. For multiple biomedical document summarization and EHRs summarization, there are only fine-tuning based methods which are proved to be effective in introducing PLMs.

Moreover, for radiology reports and medical question summarization, it shows that ensemble multiple language models can achieve the best performance [77], [86].

**What is the difference in the choice of PLMs?** The choice of PLMs has a significant influence on the performance of biomedical text summarization. For biomedical literature, it has been proven that the domain-specific language model BioBERT has better performance than BERT-base, but underperforms BERT-large [64]. PubMedBERT further outperforms BioBERT [58]. As shown in Table 5, methods using PubMedBERT which is continually pre-trained on the biomedical texts show a better performance compared with general PLMs such as BERT and RoBERTa on both PubMed-long and PubMed-short datasets. On CORD-19 and S2ORC datasets, the method GenCompareSum with SOTA PLMs T5 outperforms the method BERTSum with BERT even though T5 applied in their methods would not be fine-tuned. As for multiple document summarization, LED based on Longformer which is specially designed for long texts presents an inferior performance compared with BART, indicating that it requires more effort to address the input length limitation of PLMs in the biomedical domain

rather than directly applying solutions from the general domain.

As for EHRs, medical conversations, and medical questions, most methods use the encoder-decoder language models such as BART, T5, PEGASUS for abstractive summarization on these datasets. For EHRs, they find that domain-specific language models such as BioBERT and PubMedBERT, are not effective for radiology reports summarization, since they are pre-trained on biomedical literature [75]. Language models such as RadBERT that are pre-trained on radiology reports, are better choices [67]. For radiology reports summarization, PEGASUS achieves better performance than BART and T5 [74], [76], [86]. For medical conversation summarization, it shows that GPT-3 and BART are promising methods with limited training data [61], [80]. For medical questions, Mrini et al [87] with BART have the best performance compared with other methods based on BERT, PEGASUS, and MiniLM.

## 6.2 Limitations

Although PLMs-based methods have greatly boosted the performance of biomedical summarization, there are still several limitations.

**Developing high-quality public datasets** The development of public datasets for biomedical text summarization is imbalanced. On one hand, from the perspective of dataset types, compared with a number of public datasets in biomedical literature, there are limited released datasets for electronic health records, medical conversations, and medical questions due to privacy issues, despite the fact that there is an urgent need for developing automated text summarization methods in these texts. On the other hand, considering the task types, there are only two public datasets for multi-document summarization, while most of the existing datasets focus on single document summarization. Moreover, the size of datasets in the biomedical domain such as CHQ-Summ is generally much smaller than those in the general domain, due to the high cost and time-consuming of human annotation which additionally requires domain-specific knowledge. The lack of high-quality large-scale public datasets can hinder the development and employment of PLMs whose performance relies on the amount of data.

**Encoding long biomedical texts** PLMs have the limitation on the token length of input documents [94] due to the high time and memory consumption of attention computations in Transformer. Most PLMs-based summarization methods directly truncate input documents and only take their first 512 tokens following methods in the general domain. However, biomedical texts such as biomedical scientific papers, usually have thousands of tokens. The truncating operation losses useful information in the truncated contents of input documents. Moreover, it also leads to the loss of long-range dependencies on long biomedical documents. Although there is method [9] that investigate using PLMs that support encode long documents such as Longformer, it shows poor performance when compared with BART on biomedical texts. Therefore, it still requires more efforts to deal with the limitation to encode long biomedical texts based on PLMs efficiently.

**Incorporating domain-specific knowledge** Domain-specific knowledge is critical for understanding biomedical

cal texts. Vocabularies, taxonomies, and ontologies such as UMLS [113] are important sources of biomedical knowledge. While existing methods with PLMs is able to capture lexical knowledge in biomedical texts, they have no knowledge of words or entities that have particular domain-specific importance and their relations. Up to now, limited efforts have been proposed to incorporate external domain specific knowledge for summarization of biomedical literature and EHRs [58], [73]. It is still a limitation for existing methods on other biomedical texts to capture the knowledge of sources such as biomedical concepts, relations between concepts, and lexicographic information et al.

**Controlling factual consistency** The factual correctness of generated summaries is especially important for the real application of automatic biomedical summarizers. However, existing abstractive methods are encouraged to reconstruct gold summaries freely without word constraints on text generation. Limited attention has been focused on the factual consistency issue in the text summarization of biomedical domain. They tend to generate summaries that fabricate facts of original inputs due to freely rephrasing [102], [114], which may cause medical errors. It has shown that existing methods based on PLMs such as BART and Longformer are not able to generate factual correct summaries [9], [102]. It is still a big challenge for existing methods to control factual consistency when fine-tuning PLMs.

**Interpretability and transparency** Similar to other deep learning methods, PLMs-based methods have the well-known interpretability problem due to the black-box nature of PLMs. For users such as clinicians, they are hard to explain how and why models select specific words or sentences to yield the final summaries. If errors are consistently made by the model, it is hard for users to know why things go wrong. The explainability and transparency of models such as the inner mechanisms of their algorithms are important in constructing reliable applications for users [115]. However, no effort investigates the problem when introducing PLMs.

**Evaluations** Objective and comprehensive evaluation metrics are important to evaluate summarization methods efficiently and reliably. Most existing methods only use the ROUGE and BERTScore metrics to evaluate their models automatically similar to methods in the general domain. However, it has been reported that ROUGE and BERTScore are far from reflecting the quality of generated summaries accurately such as factual correctness, and key finding directions. Although there are efforts that explore objective human evaluation metrics for medical studies [9], [99], it still lacks an accurate automatic evaluation metric that is compatible with humans.

Overall, the performances of existing methods are still far from desirable. We believe more efforts should be proposed to address these limitations.

### 6.3 Future Directions

In this section, we further discuss promising future directions, which we hope can provide guidelines for future research.

**New large-scale public biomedical datasets** For biomedical text summarization, the annotation of the

dataset is much expensive and time-consuming than the general domain since the annotators are required to be domain experts. Moreover, for datasets such as medical conversations and questions, the privacy issue is more critical compared with biomedical literature datasets. Although there is an urgent need for summarization methods to facilitate information processing, rare attention has been paid to developing high-quality large-scale public datasets for biomedical summarization, especially for medical conversations and questions. We believe more efforts should be devoted to the development of new large-scale public biomedical datasets, i.e., unsupervised or distant-supervised automatic annotations and federated learning [116] to allow the development of models while keeping training data on the private side. Besides, more efforts should be proposed in the future to explore unsupervised, few-shot learning, and data augmentation techniques for low-resource biomedical summarization.

**Handling long biomedical documents** PLMs generally are limited to a given length of texts due to the time complexity of the model. It is an important limitation for text summarization methods since it ignores the rest content of the text, especially in the biomedical domain whose text length is relatively larger. Although there is existing research to address this issue with PLMs for long document such as Longformer, directly applying these methods in the biomedical domain is reported to have limited improvement [9]. It is urgent to investigate the effective way for PLMs in the biomedical domain to encode the full content of long texts.

**Incorporating extra knowledge** PLMs for the general domain and biomedical domain, are shown to be able to capture common sense knowledge and biomedical knowledge to a certain extent. Although they can generate summaries that are fluent or grammatically correct, it proves that most of their generated summaries are illogical or have factual errors [9], [99]. Therefore, limited knowledge captured by PLMs is hard to support the model to generate desirable summaries. It is expected that more knowledge-aware models can be proposed to incorporate extra domain-specific knowledge such as knowledge base UMLS, to improve summarization generation.

**Controllable generation** Existing methods generally yield summaries that ignore users' preferences. We believe more efforts should be developed for controlled summarization of biomedical texts, that meet the expectations and requirements of users. Methods are expected to control several attributes of generated summaries, such as length, readability, text style et al.

**Benchmarks** To facilitate the development of biomedical NLP, attempts have been made to create NLP benchmarks in the biomedical domain such as BLUE [42], GLUE [46], which include the relation extraction task, text classification task et al. However, none of the existing benchmarks includes the biomedical text summarization task. Considering the variety of types of biomedical texts including scientific papers, EHRs, conversations, questions, and categories of tasks including extractive, abstractive, and multi-documents summarization, we believe it is necessary to build a unified benchmark to support the development and fair evaluations of proposed methods.

**Multimodality** In the biomedical domain, there are rich

multimodal medical datasets such as radiology reports and associated x-rays. However, most existing methods only take biomedical texts themselves as inputs. It reports that visual features can improve the performance of text generation [117]. It expects that multimodal summarization methods can draw much attention in the future.

## 7 CONCLUSION

In this survey, we make a comprehensive overview of biomedical text summarization with large-scale pre-trained language models. We systematically review recent approaches that are based on PLMs, benchmark datasets, evaluations of the task. We categorize and compare recent approaches according to the ways they leverage PLMs and what PLMs they use. Finally, we highlight the limitations of existing methods and suggest potential directions for future research. We hope the paper can be a timely survey to help future researchers.

## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ACKNOWLEDGEMENT

This research is partially supported by the Alan Turing Institute and the Biotechnology and Biological Sciences Research Council (BBSRC), BB/P025684/1.

## AUTHOR CONTRIBUTIONS

Q Xie conducted the survey, prepared figures, and tables, and drafted the manuscript. Zh Luo, B Wang, and S Ananiadou revised the manuscript carefully. S Ananiadou supervised all the processes. All authors provided feedback and approved the final version of the manuscript.

## COMPETING INTERESTS

The authors declare no competing interests.

## REFERENCES

- [1] Z. Lu, "Pubmed and beyond: a survey of web tools for searching biomedical literature," *Database*, vol. 2011, 2011.
- [2] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.
- [3] M. Maybury, *Advances in automatic text summarization*. MIT press, 1999.
- [4] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa, and G. Del Fiol, "Text summarization in the biomedical domain: a systematic review of recent research," *Journal of biomedical informatics*, vol. 52, pp. 457–467, 2014.
- [5] P. Przybyła, A. J. Brockmeier, G. Kontonatsios, M.-A. Le Pogam, J. McNaught, E. von Elm, K. Nolan, and S. Ananiadou, "Prioritising references for systematic reviews with robotanalyst: a user study," *Research synthesis methods*, vol. 9, no. 3, pp. 470–488, 2018.
- [6] A. Esteve, A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev, and R. Socher, "Covid-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–9, 2021.
- [7] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *Journal of biomedical informatics*, vol. 42, no. 5, pp. 760–772, 2009.
- [8] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
- [9] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, "Ms<sup>2</sup>: Multi-document summarization of medical studies," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7494–7513.
- [10] B. C. Wallace, S. Saha, F. Soboczenski, and I. J. Marshall, "Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization," *AMIA Summits on Translational Science Proceedings*, vol. 2021, p. 605, 2021.
- [11] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.
- [12] J. M. Conroy and D. P. O'leary, "Text summarization via hidden markov models," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 406–407.
- [13] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 484–494.
- [14] A. Roberts, C. Raffel, and N. Shazeer, "How much knowledge can you pack into the parameters of a language model?" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5418–5426.
- [15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [16] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos, "Summarization from medical documents: a survey," *Artificial intelligence in medicine*, vol. 33, no. 2, pp. 157–177, 2005.
- [17] R. Pivovarov and N. Elhadad, "Automated methods for the summarization of electronic health records," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 938–947, 2015.
- [18] M. Wang, M. Wang, F. Yu, Y. Yang, J. Walker, and J. Mostafa, "A systematic review of automatic text summarization for biomedical literature and ehrs," *Journal of the American Medical Informatics Association*, vol. 28, no. 10, pp. 2287–2297, 2021.
- [19] B. Wang, Q. Xie, J. Pei, P. Tiwari, Z. Li et al., "Pre-trained language models in biomedical domain: A systematic survey," *arXiv preprint arXiv:2110.05006*, 2021.
- [20] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," in *Thirty-first AAAI conference on artificial intelligence*, 2017.
- [21] H. Lin and V. Ng, "Abstractive summarization: A survey of the state of the art," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9815–9822.
- [22] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 1995, pp. 68–73.
- [23] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 379–389.
- [24] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.
- [25] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.
- [26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.
- [27] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

- [28] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Jun. 2018, pp. 2227–2237.
- [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [30] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.
- [31] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1756–1765.
- [32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [34] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training,"
- [35] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.
- [36] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [37] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "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*, 2020, pp. 7871–7880.
- [38] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.
- [39] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), 2016, pp. 1715–1725.
- [40] T. Kudo and J. Richardson, "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*, 2018, pp. 66–71.
- [41] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- [42] Y. Peng, S. Yan, and Z. Lu, "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*, 2019, pp. 58–65.
- [43] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, N. Redmond, and M. B. McDermott, "Publicly available clinical bert embeddings," *NAACL HLT 2019*, p. 72, 2019.
- [44] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [45] I. Beltagy, K. Lo, and A. Cohan, "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)*, 2019, pp. 3615–3620.
- [46] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.
- [47] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A discourse-aware attention model for abstractive summarization of long documents," in *Proceedings of NAACL-HLT*, 2018, pp. 615–621.
- [48] V. Gupta, P. Bharti, P. Nokhiz, and H. Karnick, "Sumpubmed: Summarization dataset of pubmed scientific articles," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, 2021, pp. 292–303.
- [49] J. Bishop, Q. Xie, and S. Ananiadou, "Gencomparesum: a hybrid unsupervised summarization method using salience," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022, pp. 220–240.
- [50] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney *et al.*, "Cord-19: The covid-19 open research dataset," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.
- [51] Y. Guo, W. Qiu, Y. Wang, and T. Cohen, "Automated lay language summarization of biomedical scientific reviews," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 160–168.
- [52] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, pp. 1–8, 2019.
- [53] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.
- [54] Y. Song, Y. Tian, N. Wang, and F. Xia, "Summarizing medical conversations via identifying important utterances," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 717–729.
- [55] A. B. Abacha and D. Demner-Fushman, "On the summarization of consumer health questions," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2228–2234.
- [56] S. Yadav, D. Gupta, and D. Demner-Fushman, "Chq-summ: A dataset for consumer healthcare question summarization," *arXiv preprint arXiv:2206.06581*, 2022.
- [57] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang, "Extractive summarization as text matching," in *ACL*, 2020.
- [58] Q. Xie, J. A. Bishop, P. Tiwari, and S. Ananiadou, "Pre-trained language models with domain knowledge for biomedical extractive summarization," *Knowledge-Based Systems*, p. 109460, 2022.
- [59] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. S. Weld, "S2orc: The semantic scholar open research corpus," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4969–4983.
- [60] S. Enarvi, M. Amoia, M. D.-A. Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto *et al.*, "Generating medical reports from patient-doctor conversations using sequence-to-sequence models," in *Proceedings of the first workshop on natural language processing for medical conversations*, 2020, pp. 22–30.
- [61] B. Chintagunta, N. Katariya, X. Amatriain, and A. Kannan, "Medically aware gpt-3 as a data generator for medical dialogue summarization," in *Machine Learning for Healthcare Conference*. PMLR, 2021, pp. 354–372.
- [62] L. Zhang, R. Negrinho, A. Ghosh, V. Jagannathan, H. R. Hassanzadeh, T. Schaaf, and M. R. Gormley, "Leveraging pretrained models for automatic summarization of doctor-patient conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3693–3712.
- [63] W.-w. Yim and M. Yetisgen-Yildiz, "Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization," in *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, 2021, pp. 10–20.
- [64] M. Moradi, G. Dorffner, and M. Samwald, "Deep contextualized embeddings for quantifying the informative content in biomedical text summarization," *Computer methods and programs in biomedicine*, vol. 184, p. 105117, 2020.
- [65] M. Moradi, M. Dashti, and M. Samwald, "Summarization of biomedical articles using domain-specific word embeddings and graph ranking," *Journal of Biomedical Informatics*, vol. 107, p. 103452, 2020.
- [66] S. S. Gharebagh, N. Goharian, and R. Filice, "Attend to medical ontologies: Content selection for clinical abstractive summariza-

- tion," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1899–1905.
- [67] A. Yan, J. McAuley, X. Lu, J. Du, E. Y. Chang, A. Gentili, and C.-N. Hsu, "Radbert: Adapting transformer-based language models to radiology," *Radiology: Artificial Intelligence*, p. e210258, 2022.
- [68] D. Su, Y. Xu, T. Yu, F. B. Siddique, E. Barezi, and P. Fung, "Caire-covid: A question answering and query-focused multi-document summarization system for covid-19 scholarly information management," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.
- [69] J. W. Park, "Continual bert: Continual learning for adaptive extractive summarization of covid-19 literature," *arXiv preprint arXiv:2007.03405*, 2020.
- [70] Y. Du, Q. Li, L. Wang, and Y. He, "Biomedical-domain pre-trained language model for extractive summarization," *Knowledge-Based Systems*, vol. 199, p. 105964, 2020.
- [71] X. Cai, S. Liu, L. Yang, Y. Lu, J. Zhao, D. Shen, and T. Liu, "Covidsum: A linguistically enriched scibert-based summarization model for covid-19 scientific papers," *Journal of Biomedical Informatics*, vol. 127, p. 103999, 2022.
- [72] N. Kanwal and G. Rizzo, "Attention-based clinical note summarization," in *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 2022, pp. 813–820.
- [73] H. Jingpeng, L. Zhuo, C. Zhihong, L. Zhen, W. Xiang, and C. Tsung-Hui, "Graph enhanced contrastive learning for radiology findings summarization," in *Proceedings of Association for Computational Linguistics (ACL)*, vol. 2, 2022.
- [74] W. Zhu, Y. He, L. Chai, Y. Fan, Y. Ni, G. Xie, and X. Wang, "paht\_nlp@ medqa 2021: Multi-grained query focused multi-answer summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 96–102.
- [75] R. Kondadadi, S. Manchanda, J. Ngo, and R. McCormack, "Optum at medqa 2021: Abstractive summarization of radiology reports using simple bart finetuning," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 280–284.
- [76] L. Xu, Y. Zhang, L. Hong, Y. Cai, and S. Sung, "Chichealth@ medqa 2021: Exploring the limits of pre-trained seq2seq models for medical summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 263–267.
- [77] Y. He, M. Chen, and S. Huang, "damo\_nlp at medqa 2021: knowledge-based preprocessing and coverage-oriented reranking for medical question summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 112–118.
- [78] S. Yadav, D. Gupta, A. B. Abacha, and D. Demner-Fushman, "Reinforcement learning for abstractive question summarization with question-aware semantic rewards," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 249–255.
- [79] K. Krishna, S. Khosla, J. P. Bigham, and Z. C. Lipton, "Generating soap notes from doctor-patient conversations using modular summarization techniques," 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)*, 2021, pp. 4958–4972.
- [80] D. F. Navarro, M. Dras, and S. Berkovsky, "Few-shot fine-tuning sota summarization models for medical dialogues," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 2022, pp. 254–266.
- [81] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "Biobart: Pretraining and evaluation of a biomedical generative language model," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022, pp. 97–109.
- [82] A. Yalunin, D. Umerenkov, and V. Kokh, "Abstractive summarization of hospitalisation histories with transformer networks," *arXiv preprint arXiv:2204.02208*, 2022.
- [83] D. Mahajan, C.-H. Tsou, and J. J. Liang, "Ibmresearch at medqa 2021: Toward improving factual correctness of radiology report abstractive summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 302–310.
- [84] S. Yadav, D. Gupta, A. B. Abacha, and D. Demner-Fushman, "Question-aware transformer models for consumer health question summarization," *Journal of Biomedical Informatics*, vol. 128, p. 104040, 2022.
- [85] V. Kieuvoengnam, B. Tan, and Y. Niu, "Automatic text summarization of covid-19 medical research articles using bert and gpt-2," *arXiv preprint arXiv:2006.01997*, 2020.
- [86] S. Dai, Q. Wang, Y. Lyu, and Y. Zhu, "Bdkg at medqa 2021: System report for the radiology report summarization task," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 103–111.
- [87] K. Mrini, F. Démoncourt, S. Yoon, T. Bui, W. Chang, E. Farcas, and N. Nakashole, "A gradually soft multi-task and data-augmented approach to medical question understanding," 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)*, 2021, pp. 1505–1515.
- [88] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8342–8360.
- [89] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2019.
- [90] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.
- [91] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2019.
- [92] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," 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)*, 2019, pp. 3730–3740.
- [93] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.
- [94] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.
- [95] A. B. Abacha, Y. M'rabet, Y. Zhang, C. Shivade, C. Langlotz, and D. Demner-Fushman, "Overview of the medqa 2021 shared task on summarization in the medical domain," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 74–85.
- [96] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, "Big bird: Transformers for longer sequences," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 283–17 297, 2020.
- [97] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2401–2410.
- [98] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [99] J. Otmakhova, K. Verspoor, T. Baldwin, and J. H. Lau, "The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5098–5111.
- [100] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
- [101] Y. Zhang, D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz, "Learning to summarize radiology findings," in *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, 2018, pp. 204–213.
- [102] Y. Zhang, D. Merck, E. B. Tsai, C. D. Manning, and C. Langlotz, "Optimizing the factual correctness of a summary: A study of summarizing radiology reports," in *ACL*, 2020.
- [103] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. Lungren, "Combining automatic labelers and expert annotations for accurate radiology report labeling using bert," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1500–1519.
- [104] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy



- enlisted personnel," Naval Technical Training Command Millington TN Research Branch, Tech. Rep., 1975.
- [105] R. Gunning *et al.*, "Technique of clear writing," 1952.
  - [106] M. Coleman and T. L. Liao, "A computer readability formula designed for machine scoring," *Journal of Applied Psychology*, vol. 60, no. 2, p. 283, 1975.
  - [107] F. Moramarco, A. P. Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Belz, and A. Savkov, "Human evaluation and correlation with automatic metrics in consultation note generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5739–5754.
  - [108] M. Popović, "chrF: character n-gram f-score for automatic mt evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015, pp. 392–395.
  - [109] F. Moramarco, D. Juric, A. Savkov, and E. Reiter, "Towards objectively evaluating the quality of generated medical summaries," in *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, 2021, pp. 56–61.
  - [110] A. Nenkova and L. Vanderwende, "The impact of frequency on summarization," *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, vol. 101, 2005.
  - [111] M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, and Y. Yang, "LongT5: Efficient text-to-text transformer for long sequences," *arXiv preprint arXiv:2112.07916*, 2021.
  - [112] Q. Xie, J. Huang, T. Saha, and S. Ananiadou, "Gretel: Graph contrastive topic enhanced language model for long document extractive summarization," *arXiv preprint arXiv:2208.09982*, 2022.
  - [113] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl\_1, pp. D267–D270, 2004.
  - [114] W. Kryściński, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9332–9346.
  - [115] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.
  - [116] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.
  - [117] J.-B. Delbrouck, C. Zhang, and D. Rubin, "Qiai at mediq 2021: Multimodal radiology report summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 285–290.