

BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model

Hongyi Yuan¹ * Zheng Yuan¹ * Ruyi Gan² Jiaxing Zhang² Yutao Xie² Sheng Yu¹ †

¹Tsinghua University ²International Digital Economy Academy

{yuanhy20, yuanz17}@mails.tsinghua.edu.cn

{ganruyi, zhangjiaxing, xieyutao}@idea.edu.cn

syu@tsinghua.edu.cn

Abstract

Pretrained language models have served as important backbones for natural language processing. Recently, in-domain pretraining has been shown to benefit various domain-specific downstream tasks. In the biomedical domain, natural language generation (NLG) tasks are of critical importance, while under-studied. Approaching natural language understanding (NLU) tasks as NLG achieves satisfying performance in the general domain through constrained language generation or language prompting. We emphasize the lack of in-domain generative language models and the unsystematic generative downstream benchmarks in the biomedical domain, hindering the development of the research community. In this work, we introduce the generative language model BioBART that adapts BART to the biomedical domain. We collate various biomedical language generation tasks including dialogue, summarization, entity linking, and named entity recognition. BioBART pretrained on PubMed abstracts has enhanced performance compared to BART and set strong baselines on several tasks. Furthermore, we conduct ablation studies on the pretraining tasks for BioBART and find that sentence permutation has negative effects on downstream tasks.

1 Introduction

Since the advent of ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), the new pretrain-then-finetune paradigm has brought great performance improvement and dominated the methodology research of the natural language processing (NLP) field. Previous research has illustrated that pre-training language models on the domain-specific corpora can improve the model performance on domain-specific tasks further (Gururangan et al., 2020). With the large-scale publicly accessible

corpora from PubMed, researchers have already proposed biomedical domain pretrained language models such as BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2022) to aid the later research.

Natural language generation (NLG) tasks such as dialogue system (Chao et al., 2017) and question answering (Jin et al., 2022) are of critical importance for the biomedical artificial intelligence research, and there is also a trend to approach natural language understanding as NLG tasks in the general domain (Sun et al., 2021; Yan et al., 2021). For example, an entity retrieval task can be solved by constrained natural language generation (Cao et al., 2021). However, there exist two gaps in the research of the biomedical NLG. On the one hand, the architectures of the biomedical pretrained language models are almost all encoder-only transformers. Such architecture is incapable of generating natural languages auto-regressively. A decoder is necessary for language generation (Liu and Lapata, 2019). On the other hand, there are very few in-domain generative language models for biomedicine (Phan et al., 2021). Models pretrained on biomedical corpora may further enhance the performance of current biomedical NLG methods.

To bridge the gaps mentioned above, we propose a biomedical auto-regressive generative language model, BioBART, pretrained on the biomedical corpora. In our work, we adopt BART (Bidirectional and Auto-Regressive Transformers), a generative pretrained language model which achieves state-of-the-art results on different NLG tasks in the general domain (Lewis et al., 2020a). We continuously pretrain BART on PubMed abstracts to achieve biomedical domain adaption only using the text-infilling task. We also collate and evaluate BioBART on the existing biomedical NLG tasks. The in-domain BioBART outperforms BART model and sets strong baselines for several NLG tasks.

The main contributions of our work are summa-

* Contributed equally.

† Corresponded author.

rized as follows¹:

1. In aid of the research concerning the biomedical NLG tasks, we collate existing biomedical NLG tasks along with corresponding data and experimental settings. The archived biomedical tasks will be released.
2. We further analyze the influence of the pretraining task of sentence permutation in BART, and we find it brings degradation on the biomedical NLG tasks.
3. We evaluate our BioBART models on various NLG tasks and demonstrate the superb performance over BART. We will release the codes and weights to help reproduce our results.

2 Related Work

2.1 Auto-regressive Language Model

Most of the prestigious language models such as BERT, RoBERTa (Liu et al., 2019) are auto-encoding transformers. The encoder-only architecture prevents the direct implementation of the seq2seq language generation. Several generative auto-regressive language models are proposed to mitigate the problem. The serial GPT models (Radford and Narasimhan, 2018; Radford et al., 2019; Brown et al., 2020) adopt the decoder-only transformer architecture which is a left-to-right language model. They pretrain the models by auto-regressively predicting the upcoming word of sentences. UniLM1 (Dong et al., 2019) and UniLM2 (Bao et al., 2020) implement attention masks to the transformer encoder to achieve unidirectional language modeling. They pretrain their models with a mixture of masked language modeling and auto-regressive language generation. T5 (Raffel et al., 2020) and BART (Lewis et al., 2020a) apply the full transformer architecture, the encoder is used for input sequence encoding and the decoder is used for language generation. T5 and BART are both pretrained by denoising the corrupted corpora. Such models achieve many state-of-the-art results on various NLG tasks and some NLU tasks.

2.2 Biomedical Domain Pretraining

Existing work has shown that pretraining the language models on the domain-specific corpora can

bring better model transferability on the corresponding downstream tasks (Gururangan et al., 2020). There are endeavors to adapt language models to the specific domain. BioBERT (Lee et al., 2020) pretrained BERT model using biomedical corpora from PubMed abstracts and PubMed Central (PMC) full-text articles. BlueBERT (Peng et al., 2020) and clinicalBERT (Alsentzer et al., 2019) add electronic medical record (EMR) corpora from MIMIC-III (Johnson et al., 2016) to the pretraining data. Instead of continuous training from the general BERT checkpoint, SciBERT (Beltagy et al., 2019) and PubMedBERT (Gu et al., 2022) are trained from scratch using scientific papers from Semantic Scholar (Ammar et al., 2018) and PubMed articles respectively. (Shin et al., 2020) releases BioMegatron, a larger-size BERT-style language model pretrained on PubMed abstracts, PMC and MIMIC-III. The aforementioned work all use the model architecture of BERT. Other researchers are exploring different language models.

BioELMo (Jin et al., 2019) is pretrained on biomedical corpora based on stacked bidirectional LSTM language model ELMo (Peters et al., 2018). BioELECTRA (Kanakarajan et al., 2021) applies an adversarial training scheme consisting of a discriminator and a generator. They use PubMed abstracts and PMC articles as in-domain pretraining corpora. BioMed-RoBERTa (Gururangan et al., 2020) is initialized from RoBERTa (Liu et al., 2019), with additional training on the scientific papers from Semantic Scholar. Bio-lm (Lewis et al., 2020b) is pretrained on data from PubMed, PMC, and MIMIC-III based on the RoBERTa model. Ke-BioLM (Yuan et al., 2021) uses Entity as Experts (Févry et al., 2020) model to inject biomedical entity knowledge into the language model, starting from the weights of PubMedBERT. Coder (Yuan et al., 2022b) and SapBERT (Liu et al., 2021) take advantage of the synonyms resource from biomedical knowledge base UMLS (Bodenreider, 2004) and enhance the model with entity knowledge by contrastive pretraining.

Due to the nature of model architecture, encoder-only language models have limited performance on the NLG tasks, such as summarization and question answering. In recent research, SciFive (Phan et al., 2021) is proposed for biomedical NLP tasks. SciFive is pretrained on PubMed abstracts and PMC articles based on T5 architecture. While T5 is avail-

¹Our codes and pretrained checkpoints can be found at <https://github.com/GanjinZero/BioBART>.

able for NLG tasks, SciFive is focused on evaluating NLU tasks. Compared to SciFive, we choose to use BART as our model backbone and evaluate more on NLG tasks to leverage the power of decoders.

2.3 Biomedical Natural Language Generation

In the biomedical domain, most of the NLP tasks are natural language understanding (NLU) tasks. There are well-archived benchmarks for the evaluation of biomedical NLU, such as BLUE (Gu et al., 2022) and CBLUE (Zhang et al., 2021). NLG tasks are relatively less studied. (Ju et al., 2020) collects the patients and doctors’ dialogues and forms a benchmark for Covid-19 related dialogue system. (Ben Abacha et al., 2021) is an annual biomedical NLP competition containing NLG tasks such as medical question (or answer) summarization and figure captions.

Moreover, with the success of GPT-3, there is a novel trend that unifies all the NLP tasks as NLG tasks (McCann et al., 2018; Brown et al., 2020). The traditional NLU tasks can be approached by constrained language generation. Much attention is paid on the NLG methods recently. In the biomedical domain, entities are of primary concern. GENRE (Cao et al., 2021), Yuan et al. (2022a) and BARTNER (Yan et al., 2021) reach the new state-of-the-art by auto-regressive language model on entity linking and named entity recognition tasks. Such methods can be adapted to the biomedical domain.

3 Biomedical Domain Pretraining

BART is a sequence-to-sequence model with a bi-directional encoder and a left-to-right auto-regressive decoder. The model architecture is consistent with the Transformers (Vaswani et al., 2017) except for changing the ReLU activation functions to GeLUs (Hendrycks and Gimpel, 2016). BART is pretrained by denoising the corrupted input documents. The work ablates five different types of corruption noise: text masking, text deletion, text infilling, sentence permutation, and document rotation. As a result, the pretraining documents are corrupted in two ways: 1) **Text Infilling**: For each document, a number of token spans are sampled, and each sample span is replaced with a single mask token. 2) **Sentence Permutation**: A document is split into sentences and sentences are shuf-

fled in random orders. The pretraining objective is to minimize the negative log-likelihood of the original documents.

Prior work has shown that continuous-pretrained models can get competitive results compared with those trained from scratch (Gu et al., 2022). In our work, we continuously pretrain BART on the biomedical domain corpora. We revisit the methods to corrupt input texts. BART keeps the sentence permutation noise because of the significant performance gain on the summarization task, although this noise may lead to slight degradation on other tasks. We run further ablation studies on various biomedical NLG tasks. We show that the model pretrained without sentence permutation has better performance. Further details are listed in Section 5.5. Therefore we only implement the text infilling task to corrupt input texts for pretraining BioBART.

4 Generative Downstream Task

In this section, we introduce the generative downstream tasks in the biomedical domain. We will conduct experiments on these tasks to illustrate the performance of the domain-specific BioBART.

4.1 Dialogue System

A medical dialogue system aims to imitate the human doctor to communicate with human patients in a natural way. Based on the BART-style model, the patients’ primitive descriptions and dialogue histories are used as inputs to the model, then the model auto-regressively generates the replies as outputs. The task is trained and evaluated in a sequence-to-sequence fashion.

4.2 Abstractive Summarization

Summarization is a classical NLP task. It is important for healthcare to concisely summarize knowledge-rich biomedical documents. Technically, there are abstractive and extractive approaches to generate better summaries. With the help of large pretrained language models, abstractive summarization methods outperform extractive methods in summary diversity and conciseness (Zhang et al., 2020a; Dou et al., 2021). The abstractive summarization is naturally an NLG task. We follow the BART (Lewis et al., 2020a) work and evaluate our BioBART on the biomedical summarization tasks in the same fashion. The input documents are encoded by the model encoder and

the summaries are generated by the decoder auto-regressively.

4.3 Entity Linking

Entity linking is a task that maps entity mentions in texts to its standard entity concepts. Traditional entity linking methods use language models to encode entity concepts from knowledge bases(e.g. UMLS) and mentions into the same dense space and disambiguate mentions by vector similarity. The large memory footprint requirements and difficult model training hinder the development of such methods. Cao et al. (2021) proposes GENRE which uses generative language models to disambiguate entity mentions by auto-regressively generating the standard concept names conditioned on the inputs. (Yuan et al., 2022a) achieves state-of-the-art entity linking performance on various biomedical entity linking datasets by generative methods. We include this leading-edge method to show the superior performance of BioBART.

4.4 Named Entity Recognition

Named entity recognition (NER) is a critical task in the biomedical NLP community which extracts biomedical-related entities from texts. Nested and discontinuous entities widely exist in biomedical papers and EMR due to the multi-granularity semantic meanings and complex syntax structures (Yuan et al., 2020). Well-used sequential labelling framework in NER (Lample et al., 2016) is not directly fitted for nested and discontinuous NER (Finkel and Manning, 2009). Yan et al. (2021) propose BARTNER to model nested and discontinuous NER into seq2seq task by inputting sentences and outputting entities with their entity types one by one. The generative approach of BARTNER achieves state-of-the-art performance on nested and discontinuous NER datasets, and we will use it to evaluate our proposed BioBART can further enhance the performance.

5 Experiments

5.1 Pretraining

Pretraining Corpora There are two main sources of biomedical corpora: PubMed abstracts, PMC articles. In the prior work (Gu et al., 2022), training on both corpora surprisingly leads to a slight degradation in performance compared to solely training on PubMed abstracts. Therefore, we only use PubMed abstracts as the pretraining cor-

pora. The corpora contain about 41 GB of biomedical research paper abstracts on PubMed.

Pretraining Setup We continuously pretrain both large and base versions of BART for 120k steps with a batch size of 2560. We use the same vocabulary as BART to tokenize the texts. Although the input length limitation of BART is 1024, the tokenized PubMed abstracts rarely exceed 512. Therefore, for the sake of training efficiency, we truncate all the input texts to 512 maximum length. We mask 30% of the input tokens and the masked span length is determined by sampling from a Poisson distribution ($\lambda = 3$) as used in BART. We use a learning rate scheduler of 0.02 warm-up ratio and linear decay. The learning rate is set to 1e-4. We train the base version of BioBART on 2 DGX with 16 40GB A100 GPUs for about 100 hours and the large version of BioBART on the same devices for 168 hours with the help of the open-resource framework DeepSpeed (Rajbhandari et al., 2020).

5.2 Dataset for Downstream Task

5.2.1 Dialogue System

CovidDialog (Ju et al., 2020) Concerning the widespread Coronavirus disease 2019 (COVID-19) pandemic, the CovidDialog dataset is proposed to facilitate the development of dialogue system providing COVID-related consultations to people. The dataset is collected from online healthcare forums. It contains 603 consultations about COVID-19 and other related pneumonia, having 1232 utterances in total. Each consultation starts with a description related to patients' medical conditions, then followed the conversation between a doctor and a patient.

5.2.2 Abstractive Summarization

iCliniq, HealthCareMagic Both datasets are extracted from MedDialog (Zeng et al., 2020) dataset, collected from the online healthcare platform. iCliniq contains 31,062 samples and HealthCareMagic contains 226,405 samples. Each sample is comprised of a summary and corresponding dialogues between a patient and a doctor. HealthCareMagic's summaries are more abstractive and are written in a formal style, unlike iCliniq's patient-written summaries. We follow the previous work (Mrini et al., 2021) for training, developing, and testing data separations of both datasets.

MeQSum (Ben Abacha and Demner-Fushman, 2019) The dataset is created for better medical question summarization because the original patients'

Task	Dataset	Train	Dev	Test	Dataset	Train	Dev	Test	Metric
Dialogue	CovidDialog	490	63	61					Rouge,BERTscore, BLEU
Summarization	MeQSum	500	-	500	MEDIQA-ANS	38,166	174	552	Rouge, BERTscore Recall@1,@5
	iCliniq	24,851	3,105	3,108	MEDIQA-QS	1,000	50	100	
	HealthCareMagic	181,122	22,641	22,642	MEDIQA-MAS	1,104	50	80	
Entity Linking	MedMentions	122,241	40,884	40,157	NCBI	5,784	787	960	Recall@1,@5
	BC5CDR	9,285	9,515	9,654	COMETA	13,489	2,176	4,350	
	AskAPatients	16,826	1,663	1,712					
NER	ShARe13	5,146	669	5,333	ShARe14	10,380	771	7,922	Entity-level F1 score
	CADEC	4,430	898	990	GENIA	50,509	-	5,506	

Table 1: The statistics of the datasets for biomedical generative tasks. The counts for NER are entity counts.

Model	Covid19-Dialogue				
	Rouge-1	Rouge-2	Rouge-L	BLEU	BERTscore
BART BASE	27.24	12.31	25.66	10.36	0.852
BioBART BASE	28.14	12.77	26.32	11.40	0.849
BART LARGE	29.02	12.08	<u>26.93</u>	10.96	0.852
BioBART LARGE	<u>28.81</u>	13.79	26.96	12.05	<u>0.850</u>
State-of-the-art	-	-	-	7.60	-
Source	-	-	-	(Zhou et al., 2021)	-

Table 2: The main results on Dialogue System task.

questions are verbose, causing difficulty for the question-answering system. The dataset contains 1000 patients’ health questions selected from a collection distributed by the U.S. National Library of Medicine (Kilicoglu et al., 2018). Each question is annotated with a question summarization by medical experts.

MEDIQA-ANS (Savery et al., 2020) When feeling discomfort, people may turn to the internet for the answers to their medical questions. The raw searching result may be obscure for even medical experts. The dataset is proposed to emphasize the need for a medical answer summarization system in aid of better understanding biomedical materials. It consists of 156 health questions, corresponding answers to these questions, and expert-created summaries (both abstractive and extractive) of these answers. Following the paper, we use BioASQ (Tsatsaronis et al., 2015) to construct training data, MedInfo (Abacha et al., 2019) for validation, and the whole MEDIQA-ANS dataset for testing.

MEDIQA-QS, MEDIQA-MAS Both datasets are derived from the MEDIQA 2021 Tasks (Ben Abacha et al., 2021). MEDIQA-QS dataset aims to incentivize the development of new summarization approaches that address specifically the challenges of long and complex health questions. The dataset provides the validation and test sets, and MeQSum dataset is used as the training set. MEDIQA-MAS aims to prompt research that si-

multaneously aggregates and summarize the different relevant answers to a medical question. This dataset provides the validation and test sets, and MEDIQA-ANS dataset comprises the training set.

5.2.3 Entity Linking

MedMentions (Mohan and Li, 2019) MedMentions is a large-scale biomedical entity recognition dataset. The commonly used St21pv subset contains 4,392 PubMed abstracts, and over 350,000 mentions are linked to concepts of 21 selected semantic types in UMLS (Bodenreider, 2004).

BC5CDR (Li et al., 2016) BC5CDR is a benchmark for biomedical entity linking. 1500 PubMed article abstracts are annotated with 4409 chemicals, 5818 diseases entities, and 3116 chemical-disease interactions. MeSH ontology, a subset of UMLS is used to annotate entities. We follow most recent work (Angell et al., 2021; Varma et al., 2021) for data pre-processing.

NCBI (Doğan et al., 2014) The dataset is built from 793 PubMed abstracts. It consists of 6892 annotated disease mentions of 790 unique disease concepts. The annotators label all the mentions to concepts in MEDIC ontology (Davis et al., 2012). MEDIC is a medical dictionary that merges the diseases concepts, synonyms, and definitions in MeSH and OMIM and is composed of 9700 unique diseases. We follow BioSyn (Sung et al., 2020) to process data and construct dataset splits.

Model	iCliniq		HealthCareMagic		MEDIQA-QS	
	Rouge-1/2/L	BERTscore	Rouge-1/2/L	BERTscore	Rouge-1/2/L	BERTscore
BART BASE	61.43/48.68/ 59.71	0.941	46.81/26.19/44.34	0.918	28.82/10.99/26.99	0.896
BioBART BASE	61.07/48.47/59.42	0.941	46.67/26.03/44.11	0.918	30.12/11.28/27.44	0.898
BART LARGE	59.87/47.01/58.12	0.938	47.24/26.54/44.68	0.919	29.97/10.64/28.41	0.901
BioBART LARGE	60.32/47.98/58.69	0.940	46.54/26.14/44.23	0.919	31.97/12.39/29.70	0.903
State-of-the-art	62.3/48.7/58.5	-	46.9/24.8/43.2	-	35.14/16.08/31.31	-
Source	(Mrini et al., 2021)		(Mrini et al., 2021)		(Ben Abacha et al., 2021)	
MEDIQA-MAS						
Model	Rouge-1/2/L		BERTscore		Rouge-1/2/L	
	31.63/9.98/27.85	0.859	19.10/6.77/16.90	0.851	52.93/35.79/50.46	0.927
BART BASE	32.90/11.28/29.26	0.861	18.97/7.46/16.77	0.850	53.75/36.50/51.27	0.929
BioBART BASE	29.32/9.00/26.14	0.857	21.52/9.31/19.15	0.853	53.68/36.80/51.05	0.928
BART LARGE	30.60/10.37/27.04	0.861	21.58/9.34/19.18	0.857	55.61/38.11/53.15	0.933
BioBART LARGE	32.15/ 16.21 /19.10	-	23.07 /5.41/15.35	-	54.5/37.9/50.2	-
State-of-the-art	74.6/-	-	91.9/-	92.4 /-	80.1/-	89.0/-
Source	(Varma et al., 2021)		(Varma et al., 2021)	(Lai et al., 2021)	(Lai et al., 2021)	(Liu et al., 2021)

Table 3: The main results on Summarization tasks.

Model	MedMentions	BC5CDR	NCBI	COMETA	AAP
	Recall@1/@5	Recall@1/@5	Recall@1/@5	Recall@1/@5	Recall@1/@5
BART BASE	69.77/84.59	91.56/94.89	88.54/95.31	78.34/87.40	86.37/94.29
BioBART BASE	71.15/86.22	93.01/95.59	89.27/95.31	79.63/88.64	87.51/94.92
BART LARGE	71.49/84.95	92.48/95.26	90.21/95.52	80.70/88.65	88.79/96.59
BioBART LARGE	71.78/85.42	93.26/95.74	89.90/95.63	81.77/88.87	89.40/95.76
State-of-the-art	74.6 / -	91.9/-	92.4 / -	80.1/-	89.0/-
Source	(Varma et al., 2021)	(Varma et al., 2021)	(Lai et al., 2021)	(Lai et al., 2021)	(Liu et al., 2021)

Table 4: The main results on Entity Linking tasks.

Model	ShARe13	ShARe14	CADEC	GENIA
	F1	F1	F1	F1
BART BASE	76.63	77.87	68.37	78.06
BioBART BASE	78.78	79.17	68.39	78.43
BART LARGE	79.69	80.34	70.64	78.93
BioBART LARGE	80.75	80.41	70.53	79.93
State-of-the-art	82.52	81.75	73.21	81.39
Source	(Li et al., 2021)			

Table 5: The main result on NER tasks.

COMETA (Basaldella et al., 2020) COMETA is derived from the online publicly available and anonymous health discussion on Reddit. It consists of 20k English biomedical entity mentions expert-annotated with concepts from SNOMED CT. We use the “stratified (general)” split and follow the training and evaluation procedures of SapBert (Liu et al., 2021) and ResCNN (Lai et al., 2021).

AskAPatient (Limsopatham and Collier, 2016) It contains 8,662 phrases from social media. Each phrase can be mapped to one of the 1,036 medical concepts from SNOMED-CT and AMT (the Australian Medicines Terminology). The samples in AskAPatient do not include contextual information. We follow Sung et al. (2020) and Limsopatham and Collier (2016) for data pre-processing and apply the 10-fold evaluation protocol.

5.2.4 Named Entity Recognition

ShARe13, ShARe14, CADEC These three datasets annotate discontinuous adverse drug events entities. The main difference is the annotated data of ShARe tasks (Pradhan et al., 2013; Mowery et al., 2014) comes from MIMIC-II, and CADEC (Karimi et al., 2015) comes from social media. There is only one entity type for these datasets. We follow Yan et al. (2021) for dataset preprocess.

GENIA (Kim et al., 2003) GENIA annotates 2000 MEDLINE abstracts with biological entities. Entities can be nested with others. We follow (Lin et al., 2019) to combine fine-grained entity types into 5 coarse-grained entity types and to construct dataset splits.

All the aforementioned datasets are in English. The statistical overview of the aforementioned datasets is listed in Table 1.

5.3 Fine-tuning details

Dialogue We use BioBART as the dialogue system model. The dialogue history is fed into the encoder and the decoder generates the response autoregressively. We apply the negative log-likelihood function as the training objective with respect to

the reference dialogue response. We fine-tune the model with learning rate 5e-5 for the base version and 1e-5 for the large version for 20 epochs. We run evaluations on the validation set at the end of each epoch and use the checkpoint with the best validation performance for testing. During inference, we use beam search of size 5 to sample responses from the model’s outputs. We use Rouge-1/2/L (Lin, 2004), BLEU (Papineni et al., 2002) and BERTscore (Zhang et al., 2020b) as our evaluation metrics. RoBERTa-large (Liu et al., 2019) is used as scorer in BERTscore.

Summarization Similarly, for summarization, the encoder takes the documents as input, and the decoder generates the corresponding summarizations. We minimize the log-likelihood objective to fine-tune the model and apply beam search for inference. Across different summarization datasets, the beam size is set to 5 and we use no length penalty. We fine-tune the model with learning rate 5e-5 for the base version and 1e-5 for the large version for 6 epochs. We run evaluations on the validation set at the end of each epoch and use the checkpoint with the best validation performance for testing. We apply the commonly used Rouge-1/2/L and BERTscore for evaluation metrics. The large version of RoBERTa is used as the scorer in BERTscore.

Entity Linking We follow the method and experimental settings in Yuan et al. (2022a) to implement the generative model for biomedical entity linking tasks. We do not apply knowledge-base guided pre-training proposed in Yuan et al. (2022a). The documents with the positions of mentions marked are fed into the encoder and the decoder outputs the corresponding synonyms in the knowledge base directly. We use the top1 and top5 recall (Recall@1 and Recall@5) as the evaluation metrics.

NER We use BARTNER (Yan et al., 2021) as our model. The target type for BARTNER is *word* (i.e. output first BPE of each word in entities). We use the parameters selected by Yan et al. (2021) for all pretrained models and fine-tune for 30 epochs. Entity-level F1 is used as the metric.

5.4 Main Result

In this section, we present the base and large version of BioBART on various generation tasks. We compare our in-domain BioBART with BART to illustrate the effectiveness of domain adaption. We

also compare with the existing state-of-the-art results on each dataset to shed light on the superior performance of BioBART. The experimental results are shown in Table 2-5. The best and the second-best scores are highlighted with bold numbers and underlines respectively.

Dialogue We evaluate biomedical dialogue response generation on CovidDialog. For both base and large version, BioBART shows improvement on the automatic metric Rouge. The large BioBART outperforms BART by 1.71 on Rouge-2 and 0.03 on Rouge-L . Our evaluations surpasses the current state-of-the-art on BLEU score by 4.45.

Summarization We present broad experimental results on biomedical summarization datasets. From Table 3, BioBART has competitive or even superior performance on the task. Except for iCliniq and HealthCareMagic, we see consistent improvement on different datasets for both sizes of BioBART. For MeQSum, BioBART large exceeds BART large for 1.93/1.31/2.1 on Rouge-1/2/L and even outperforms the current state-of-the-art. The possible reason that biomedical in-domain pretraining fails on iCliniq and HealthCareMagic is that both datasets are built upon a clinical corpus. There still exists a domain-shifting problem for BioBART pretrained on biomedical scientific articles from PubMed.

On dialogue and summarization tasks, there are minor changes in BERTscore for different models. This is possible because the metric is calculated by other pretrained language models. The implemented RoBERTa may suffer from biomedical domain-shifting and cannot quantify the model performance accurately.

Entity Linking The results on biomedical entity linking tasks are shown in Table 4. For all the tasks, models finetuned based on BioBART have better performance. On AAP, BC5CDR, and COMETA, our results outperform the current discriminative state-of-the-art methods by 0.4, 1.67, and 1.36 points of Recall@1 respectively.

NER The performance improvement of BioBART on ShAre13, ShAre14, and GENIA is significant, while the increase on CADEC is mediocre. For the large models, BioBART improves entity-level F1 scores for 1.06 and 1 on ShAre13 and GENIA datasets. There are promising results for generative biomedical NER methods, while the gap

	CovidDialogue		MeQSum		MEDIQA-MAS	
	Rouge-2/L	BLEU	Rouge-2/L	BERTscore	Rouge-2/L	BERTscore
BART BASE	12.31/25.66	10.36	35.79/50.46	0.927	9.98/27.85	0.859
w/ TI & SP	10.90/25.46	10.46	34.93/49.28	0.926	9.04/27.43	0.859
w/ TI	11.81/25.79	12.79	37.14/51.71	0.929	10.66/27.65	0.862
MedMentions		COMETA		ShARe13	CADEC	
Recall@1		Recall@1	Recall@5	F1	F1	
BART BASE	69.77	84.59	78.34	87.40	76.63	68.37
w/ TI & SP	70.76	85.69	79.31	87.93	77.01	67.98
w/ TI	70.82	85.68	79.43	88.28	78.03	69.22

Table 6: Ablation study for pretraining tasks. TI is short for Text Infilling and SP is short for Sentence Permutation.

Context(Document)	BART Response(Summary)	BioBART Response(Summary)
Description: I have type 1 diabetes that I try hard to control, but i struggle with. Are my chances high for poor outcomes with the virus? I'm in such a panic about it.	In brief: diabetes. Your risk of complications with covid-19 is very low. If you are well controlled your risk should not be elevated. If your diabetes is uncontrolled you should not have a particularly high risk.	Diabetes and Covid-19. If you are a type 1 diabetic, you are at high risk for poor outcomes with the virus. Your risk is very high. You should be aware of your risk factors and monitor your glycemic control closely.
SUBJECT: treatment for dry mouth caused by necessary medicine MESSAGE: My provider can't help (I asked.) I am intolerant of all the "sugar alcohols" such as maltitol, ...	What are the treatments for dry mouth caused by sugar alcohols?	What are the treatments for dry mouth?
Are there any new treatments for orthostatic tremor? ... Deep brain stimulation is sometimes done when drugs cannot control a severe, disabling tremor. Sometimes essential tremors or tremors due to Parkinson disease or another disorder require such treatment. ... There is no cure for most forms of tremor, but there are treatments to help manage symptoms. ...	There is no cure for tremor, but there are treatments to help manage the symptoms.	There is no cure for tremor, but there are treatments to help manage the symptoms. The most common ... help relieve symptoms. Deep brain stimulation is sometimes done when drugs cannot control a severe, disabling tremor. Sometimes essential tremors or tremors due to Parkinson disease or another disorder require such treatment.

Table 7: Example dialogue and summaries from the fine-tuned BART and BioBART model excepts of showing the references. The key information and differences are highlighted with colors.

with the current state-of-the-art NER method (Li et al., 2021) is still salient.

5.5 Ablation Study on Pretraining Task

In this section, we test on pretraining with or without the sentence permutation task. We pretrain BART base following the same pretraining settings except for reducing the training step to 40k for efficiency. We fine-tuned the pretrained models on the downstream tasks. The ablation results are shown in Table 6.

From the result, it is illustrated that the model pretrained on isolated text infilling task performs the best. The sentence permutation task downgrades the model’s performance even for generative summarization and dialogue system tasks.

5.6 Generated example

Here we demonstrate BioBART’s performance qualitatively. In Table 7, we present three generative examples on CovidDialog, MeQSum, and MEDIQA-ANS respectively. In the first example,

we can see that BART generates an erroneous instruction of the influence of diabetes. BioBART injected with domain knowledge can correctly give the response. In the second, BART misunderstands the document where sugar alcohol is not the cause of dry mouth. BioBART generates an accurate and concise summary. In the final example, the MEDIQA-ANS document is rather long and BART fails to extract complete information (colored in red). From the examples, we can conclude that BioBART has improvements on biomedical common sense and documents understanding.

6 Conclusions

In this work, we pretrain the biomedical domain generative language model BioBART. We also collect various publicly available benchmarks for biomedical generative tasks to prompt future research. Our experimental results show that continuous pretraining on PubMed abstracts helps the model with domain adaption. BioBART shows

great improvements on different benchmarks and achieves competitive or superior results over the current state-of-the-art methods. We also release our pretraining and fine-tuning codes to facilitate future research for reproducibility.

We will explore pretraining generative language models 1) on in-domain vocabularies and from scratch, 2) and with clinical corpora such as EMRs in MIMIC-III (Johnson et al., 2016) or PMC-Patients (Zhao et al., 2022) in the future studies.

Acknowledgements

We appreciate three anonymous reviewers for helpful comments. This work was supported by the National Natural Science Foundation of China (Grant No. 12171270), and the Natural Science Foundation of Beijing Municipality (Grant No. Z190024).

References

- Asma Ben Abacha, Yassine Mrabet, Mark E. Sharp, Travis R. Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers’ medication questions and trusted answers. *Studies in health technology and informatics*, 264:25–29.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.
- Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu A. Ha, Rodney Michael Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler C. Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna L. Power, Sam Skjonsberg, Lucy Lu Wang, Christopher Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *NAACL*.
- Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. 2021. Clustering-based inference for biomedical entity linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2598–2608, Online. Association for Computational Linguistics.
- Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Song-hao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR.
- Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*.
- Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28th - August 2*.
- Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDICA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online. Association for Computational Linguistics.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.
- Hsiao-Tuan Chao, Lucy Liu, and Hugo J Bellen. 2017. Building dialogues between clinical and biomedical research through cross-species collaborations. In *Seminars in cell & developmental biology*, volume 70, pages 49–57. Elsevier.
- Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. 2012. Medic: a practical disease vocabulary used at the comparative toxicogenomics database. *Database*, 2012.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

- for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.
- Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. **GSum: A general framework for guided neural abstractive summarization**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.
- Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4937–4951, Online. Association for Computational Linguistics.
- Jenny Rose Finkel and Christopher D Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 141–150.
- Yuxian Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.
- Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89.
- Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Moshia Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: A survey of approaches and challenges. *ACM Comput. Surv.*, 55(2).
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.
- Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, and Pengtao Xie. 2020. Coviddialog: Medical dialogue datasets about covid-19. <https://github.com/UCSD-AI4H/COVID-Dialogue>.
- Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: pretrained biomedical text encoder using discriminators. In *BIONLP*.
- Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.
- Halil Kılıcoglu, Asma Ben Abacha, Yassine Mrabet, Sonya E. Shooshan, Laritza M. Rodriguez, Kate Masterton, and Dina Demner-Fushman. 2018. Semantic annotation of consumer health questions. *BMC Bioinformatics*, 19.
- J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.
- Tuan Lai, Heng Ji, and ChengXiang Zhai. 2021. BERT might be overkill: A tiny but effective biomedical entity linker based on residual convolutional neural networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1631–1639, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2021. Domain adaptation with pre-trained transformers for query focused abstractive text summarization. *arXiv preprint arXiv:2112.11670*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained

- biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. **BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020b. **Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art**. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.
- Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.
- Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2021. Unified named entity recognition as word-word relation classification. *arXiv preprint arXiv:2112.10070*.
- Nut Limsopatham and Nigel Collier. 2016. **Normalising medical concepts in social media texts by learning semantic representation**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2019. **Sequence-to-nuggets: Nested entity mention detection via anchor-region networks**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5182–5192, Florence, Italy. Association for Computational Linguistics.
- Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.
- Yang Liu and Mirella Lapata. 2019. **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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.
- Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.
- Sunil Mohan and Donghui Li. 2019. **Medmentions: A large biomedical corpus annotated with {umls} concepts**. In *Automated Knowledge Base Construction (AKBC)*.
- Danielle L Mowery, Sumithra Velupillai, Brett R South, Lee Christensen, David Martinez, Liadh Kelly, Loraine Goeuriot, Noemie Elhadad, Sameer Pradhan, Guergana Savova, et al. 2014. Task 2: Share/clef ehealth evaluation lab 2014. In *Proceedings of CLEF 2014*.
- Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021. **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)*, pages 1505–1515, Online. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: A method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, page 311–318, USA. Association for Computational Linguistics.
- Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on bert for biomedical text mining. In *BIONLP*.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.
- Long Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltzman, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *ArXiv*, abs/2106.03598.

- Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee M Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana K Savova. 2013. Task 1: Share/clef ehealth evaluation lab 2013. In *CLEF (Working Notes)*, pages 212–31.
- Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC ’20. IEEE Press.
- Max E. Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7.
- Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, Online. Association for Computational Linguistics.
- Tianxiang Sun, Xiangyang Liu, Xipeng Qiu, and Xuanjing Huang. 2021. Paradigm shift in natural language processing. *arXiv preprint arXiv:2109.12575*.
- Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *ACL*.
- George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Palouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16.
- Maya Varma, Laurel Orr, Sen Wu, Megan Leszczynski, Xiao Ling, and Christopher Ré. 2021. Cross-domain data integration for named entity disambiguation in biomedical text. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4566–4575, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822, Online. Association for Computational Linguistics.
- Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022a. Generative biomedical entity disambiguation via knowledge base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *BIONLP*.
- Zheng Yuan, Yuanhao Liu, Qiuyang Yin, Boyao Li, Xiaobin Feng, Guoming Zhang, and Sheng Yu. 2020. Unsupervised multi-granular chinese word segmentation and term discovery via graph partition. *Journal of Biomedical Informatics*, 110:103542.
- Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022b. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, page 103983.
- Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. **Bertscore: Evaluating text generation with bert**. In *International Conference on Learning Representations*.

Zhengyun Zhao, Qiao Jin, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient notes and relations extracted from case reports in pubmed central. *arXiv preprint arXiv:2202.13876*.

Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, Yichen Zhang, Qingyang Wu, Zhou Yu, Kun Xu, Eric Xing, and Pengtao Xie. 2021. **On the generation of medical dialogs for COVID-19**. 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)*, pages 886–896, Online. Association for Computational Linguistics.