ClinicalT5: A Generative Language Model for Clinical Text

Qiuhao Lu[1], **Dejing Dou**[1,2], **and Thien Huu Nguyen**[1]
[1]Dept. of Computer Science, University of Oregon, Eugene, OR, USA
[2]Baidu Research
{luqh, dou, thien}@cs.uoregon.edu

Abstract

In the past few years, large pre-trained language models (PLMs) have been widely adopted in different areas and have made fundamental improvements over a variety of downstream tasks in natural language processing (NLP). Meanwhile, domain-specific variants of PLMs are being proposed to address the needs of domains that demonstrate a specific pattern of writing and vocabulary, e.g., BioBERT for the biomedical domain and ClinicalBERT for the clinical domain. Recently, generative language models like BART and T5 are gaining popularity with their competitive performance on text generation as well as on tasks cast as generative problems. However, in the clinical domain, such domain-specific generative variants are still underexplored. To address this need, our work introduces a T5-based text-to-text transformer model pre-trained on clinical text, i.e., ClinicalT5. We evaluate the proposed model both intrinsically and extrinsically over a diverse set of tasks across multiple datasets, and show that ClinicalT5 dramatically outperforms T5 in the domain-specific tasks and compares favorably with its close baselines.[1]

1 Introduction

In the past few years, large pre-trained language models (PLMs), such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020), BART (Lewis et al., 2020), T5 (Raffel et al., 2020), etc., have achieved great success over a variety of downstream tasks in natural language processing (NLP). These PLMs mainly depend on self-supervised pre-training on large amounts of general-domain textual data, e.g., Wikipedia, news articles, web crawl corpus, etc., and are widely adopted in downstream applications. Despite the superior performance of these PLMs on general-domain text, their performance over domain-specific text is relatively poor (Ma et al., 2019). To bridge this gap, researchers propose to build domain-specific PLMs through fine-tuning or pre-training from scratch over domain corpora. For example, in the biomedical and clinical domains, various domain-specific PLMs have been explored and released, including BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), BlueBERT (Peng et al., 2019), ClinicalBERT (Huang et al., 2019), BioClinicalBERT[2] (Alsentzer et al., 2019), umlsBERT (Michalopoulos et al., 2020), diseaseBERT (He et al., 2020), SciFive (Phan et al., 2021), and BioBART (Yuan et al., 2022).

Domain-specific language models have been extensively explored in different kinds of NLP-related downstream applications, ranging from entity linking (Bhowmik et al., 2021) to document classification (Allada et al., 2021). Generally, a typical and popular usage of the aforementioned PLMs is to leverage them to encode domain text, the learned representations of which are then fed into some task-specific structures for label prediction. Taking a complicated real-world task as an example, (Huang et al., 2019) predicts patients' risk of readmission within 30 days after discharge using clinical notes in the Electronic Health Records (EHRs). Essentially, they encode discharge summaries of patients with ClinicalBERT, and put the learned embeddings of the `[CLS]` token to a linear layer on top for prediction, leading to better performance than traditional models. Moreover, (Lu et al., 2021c) constructs a document-level multi-view graph out of each clinical note and predicts patients' 30-day readmission risk with a graph-based model, and they use BioClinicalBERT (Alsentzer et al., 2019) as the encoder within the graph model.

Recently, generative language models, e.g., BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), have attracted attention since they are naturally effective for natural language generation tasks, such as document summarization (Chen and

[1]We will release the models upon decision of the paper.

[2]Also known as ClinicalBERT.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5465 - 5472

Yang, 2021), question answering (Zhu et al., 2021; Sachan et al., 2021), data augmentation (Lu et al., 2021b), etc. Meanwhile, a novel paradigm of leveraging generative language models has gained popularity, where researchers cast non-generation tasks as generative problems, e.g., to directly generate textual labels to incorporate their semantics, and report promising results (De Cao et al., 2021; De Cao et al., 2022). However, such approaches are still underexplored in certain domains due to lack of domain-specific generative language models, i.e., most of the aforementioned domain-specific PLMs are notably domain-adapted BERT-style models. In the biomedical domain, two generative language models SciFive (Phan et al., 2021) and BioBART (Yuan et al., 2022) have been released, but in the clinical domain, the situation is worse and no such generative models exist to our knowledge. Though the two domains are relatively close, clinical text poses unique challenges compared to general and non-clinical biomedical text due to its specific linguistic characteristics (Alsentzer et al., 2019). Previous studies list some of the linguistic features of clinical text, e.g., heavy use of professional technical terminology, abbreviations and acronyms, passive verbs, omission of subjects and verbs, etc., and these features make clinical text divergent from standard language (Smith et al., 2014).

Aiming to fulfill this gap, we adapt T5 (Raffel et al., 2020) to the clinical domain by training a domain-specific variant using clinical text, i.e., ClinicalT5. We demonstrate the capabilities of the model by conducting both intrinsic and extrinsic evaluations. For intrinsic evaluation, we aim to evaluate its capability to capture the similarity and relatedness of the Unified Medical Language System (UMLS) concept pairs, where we measure the correlation coefficient between the similarity scores of the encoded representations for the concept pairs and those judged by human experts. For extrinsic evaluation, we evaluate the proposed model along with baselines over a diverse set of benchmark datasets, ranging from document classification (DC), named entity recognition (NER), to natural language inference (NLI). Furthermore, we also evaluate on three more complicated real-world tasks of clinical importance, i.e., patients' 30-day readmission risk, 30-day and 1-year mortality risk. We show that ClinicalT5 dramatically outperforms T5 and compares favorably with its close baselines across all of these tasks.

2 Related Work

2.1 Biomedical Domain-Adapted Models

The biomedical domain has been an active area of research in the NLP community for the past few years. Many relevant studies have been presented, ranging from domain-specific language models, external knowledge infusion, and various downstream applications, etc. (Peng et al., 2019; Beltagy et al., 2019; Lee et al., 2020; He et al., 2020; Michalopoulos et al., 2020; Lu et al., 2021a). Most of the biomedical language models are BERT (Devlin et al., 2019) variants fine-tuned to biomedical text, e.g., BioBERT is trained on PubMed abstracts and PMC full text articles (Lee et al., 2020) and SciBERT is trained on the full text of biomedical and computer science papers from the Semantic Scholar corpus (Beltagy et al., 2019). Besides, researchers inject external domain knowledge into adapted biomedical language models due to the knowledge-intensive nature of this domain, e.g., umlsBERT is directly trained using UMLS text (Michalopoulos et al., 2020), He *et al.* infuse disease information from the corresponding Wikipedia passages into language models (He et al., 2020), and Lu *et al.* inject biomedical knowledge from multiple sources into language models via adapters (Lu et al., 2021a). For generative language models, SciFive is an adapted T5 model pre-trained on PubMed abstracts and PMC articles (Phan et al., 2021) and BioBART is an adapted BART model pre-trained on PubMed abstracts (Yuan et al., 2022).

2.2 Clinical Domain-Adapted Models

In the clinical domain, there are mainly two popular BERT models, i.e., ClinicalBERT (Huang et al., 2019) and BioClinicalBERT (Alsentzer et al., 2019), which are both trained on the clinical notes in the MIMIC-III database (Johnson et al., 2016). For generative language models, however, the topic is not well explored and this situation motivates our work.

3 ClinicalT5

Following prior studies on clinical language models (Huang et al., 2019; Alsentzer et al., 2019), we use the textual notes in MIMIC-III to train ClinicalT5, which consists of approximately 2 million notes. Similarly, only minimal pre-processing is conducted where unnecessary tokens and characters are removed (Huang et al., 2019).

In particular, we initialize the weights from the SciFive-PubMed-PMC model (base and large) (Phan et al., 2021) and further pre-train with the span-mask denoising objective (Raffel et al., 2020) on the pre-processed MIMIC-III notes. The base and large models have $\sim 220M$ parameters with 12 layers and $\sim 770M$ parameters with 24 layers, respectively. For each of the two versions, we further pre-train ClinicalT5 on the unlabeled text for extra $10k$ steps, with a max sequence length of 512, a batch size of 8, and a learning rate of $1e{-}4$. The pre-training is performed on 3 Nvidia Tesla V100-32GB GPUs. We provide a reproducibility checklist in Appendix A, and we refer the readers to (Raffel et al., 2020) for more detailed treatment of the architecture and training objectives of T5.

4 Experiments

In this section, we evaluate ClinicalT5 both intrinsically and extrinsically, along with the following generative baselines (for both general and domain-specific texts): BART (Lewis et al., 2020), BioBART (Yuan et al., 2022), T5 (Raffel et al., 2020), SciFive (Phan et al., 2021), to demonstrate the capabilities of ClinicalT5 across different applications.

4.1 Intrinsic Evaluation

We conduct intrinsic evaluation on the datasets UMNSRS-Sim and UMNSRS-Rel (Pakhomov et al., 2010), which consist of 566 and 587 UMLS term pairs respectively. Each pair comes with a *similarity* score and a *relatedness* score that are manually assigned by human experts. Similar to previous work (Zhang et al., 2019), we encode the terms with ClinicalT5 and the baselines. Essentially, we use the mean-pooled vectors of the last hidden states of the encoders as the term embeddings and calculate a cosine similarity score for each pair. Then we compute the Pearson's correlation coefficient and Spearman's correlation coefficient between the computed scores and the expert-assigned scores. As shown in Table 1, ClinicalT5 demonstrates a better ability to capture the similarity of UMLS terms than T5 and Scifive, indicating the effectiveness of the training.

4.2 Extrinsic Evaluation

For extrinsic evaluation, we consider three different tasks, i.e., document classification (DC), named entity recognition (NER), and natural language inference (NLI). To validate the models' capability on clinical text, we select datasets that are closely relevant to clinical targets rather than biomedical or chemical related data such as BC5CDR-chemical (Li et al., 2016). We fine-tune the evaluating models on 4 corresponding datasets across these tasks in a single-task text-to-text manner. For all the experiments, we use a batch size of 16 and a learning rate of $1e{-}4$. Due to different targets, we set the max source text length to 256, and the max target text lengths to 52, 256, 256, 15 for the datasets HOC, NCBI, BC5CDR and MEDNLI, respectively.

Model	UMNSRS-Similarity		UMNSRS-Relatedness	
	Pearson	Spearman	Pearson	Spearman
BART-base	0.1456	0.1300	0.0756	0.0625
BioBART-base	**0.3753**	**0.3441**	**0.3101**	**0.2929**
T5-base	0.2050	0.1448	0.1056	0.0519
SciFive-base	0.1941	0.1488	0.1359	0.0900
ClinicalT5-base	0.2126	0.1611	0.1478	0.0948
BART-large	0.2234	0.1958	0.1706	0.1546
BioBART-large	**0.4511**	**0.4302**	**0.3517**	**0.3400**
T5-large	0.2379	0.2018	0.1813	0.1564
SciFive-large	0.3176	0.2642	0.3039	0.2618
ClinicalT5-large	0.3391	0.2847	0.2884	0.2468

Table 1: Pearson's and Spearman's correlation coefficient scores.

4.2.1 Document Classification

We conduct document classification on the HOC dataset (Baker et al., 2016), which consists of $9,972$ samples for training and $4,947$ samples for testing. Essentially, we fine-tune the evaluating models to categorize the texts into 10 categories by directly generating the class labels, e.g., "`empty`", "`evading growth suppressors`", "`genomic instability and mutation`", etc.

4.2.2 Named Entity Recognition

We conduct named entity recognition on two popular datasets, i.e., NCBI-disease (Doğan et al., 2014) and BC5CDR-disease (Li et al., 2016). The input text sequence may contain a disease term and the term should be identified and labeled in the target text, e.g., for the input text "Genotype and phenotype in patients with dihydropyrimidine dehydrogenase deficiency", the target is "Genotype and phenotype in patients with `disease*` dihydropyrimidine dehydrogenase deficiency `*disease`".

4.2.3 Natural Language Inference

We conduct natural language inference evaluation on the MEDNLI dataset (Romanov and Shivade, 2018), which consists of $11,232$ training samples

Tasks	HOC			NCBI			BC5CDR			MEDNLI
Metrics(%)	P	R	F1	P	R	F1	P	R	F1	Acc
BART-base	80.30	79.84	79.81	62.23	72.09	66.80	59.24	67.26	63.00	75.60
BioBART-base	84.68	83.54	83.82	63.10	71.77	67.16	61.78	72.05	66.52	80.66
T5-base	82.00	80.98	81.19	86.64	83.00	84.78	80.73	81.68	81.20	81.86
SciFive-base	85.10	84.83	84.70	86.43	88.25	87.33	**83.56**	81.43	**82.48**	83.90
ClinicalT5-base	**85.44**	**85.14**	**85.06**	**87.28**	**88.56**	**87.92**	81.55	**82.92**	82.23	**84.95**
BART-large	84.89	84.07	84.18	63.39	74.50	68.50	66.45	62.07	64.19	84.53
BioBART-large	84.80	84.51	84.39	67.74	70.51	69.10	65.00	71.93	68.29	**86.29**
T5-large	85.42	84.75	84.79	84.20	84.99	84.60	78.31	79.75	79.02	83.83
SciFive-large	**85.57**	**85.67**	**85.34**	85.91	85.10	85.50	78.28	79.89	79.08	84.95
ClinicalT5-large	85.37	84.79	84.78	**86.37**	**87.09**	**86.73**	**79.24**	**81.49**	**80.35**	85.86

Table 2: Performance comparison over document classification, named entity recognition, and medical natural language inference.

and $1,422$ testing samples. Essentially, we convert the premise-hypothesis pair to a sequence and prepend a task-specific prefix to it, e.g., "mednli: premise: [...]. hypothesis: [...]." We take the converted sequence as the input text and fine-tune the evaluating models to generate the target labels, i.e., "`contradiction`", "`neutral`", "`entailment`".

4.2.4 Results

The results are shown in Table 2. Generally, ClinicalT5 outperforms T5 and SciFive across most of these metrics, and the advantage indicates the success of the training over clinical text. However, ClinicalT5-large is on par with T5-large and has a slightly lower recall than SciFive-large on the HOC dataset. We conjecture that the large versions of BART and T5 already have enough capacity for the task which makes domain-specific training less impressive, as reflected by the fact that BioBART-large is only marginally better than BART-large. For MEDNLI, ClinicalT5 consistently outperforms T5 and SciFive although BioBART-large achieves the highest accuracy.

4.3 Real-world Evaluation

We also evaluate the models on more complicated real-world applications of clinical importance, i.e., 30-day unplanned ICU patient readmission risk, 30-day and 1-year patient mortality risk. The experiment is conducted based on the MIMIC-III dataset (Johnson et al., 2016). Following previous work (Zhang et al., 2020; Lu et al., 2021c), we extract the discharge summaries from EHRs and generate $48,393$ documents. Essentially, we take the evaluating models to encode the last 512 tokens of each note, the last hidden states of which are fed into a linear layer on top for prediction. As shown in Table 3, ClinicalT5 shows the best results across almost all the metrics, demonstrating its potential for real-world applications in the clinical domain.

Tasks	30-d Readmission			30-d Mortality		1-y Mortality	
Metrics(%)	A.R.	A.P.	RP80	A.R.	A.P.	A.R.	A.P.
T5-base	77.10	52.24	16.97	80.03	23.62	78.52	45.72
SciFive-base	**78.12**	53.95	18.87	80.38	24.16	78.95	45.38
ClinicalT5-base	77.94	**54.25**	**19.76**	**81.11**	**26.70**	**79.09**	**46.58**

A.R: AUC under ROC, A.P: AUC under PRC, RP80: recall at precision of 80%

Table 3: Performance on patients' outcomes prediction.

5 Conclusion

In this study, we explore and propose ClinicalT5, a T5-based text-to-text transformer model for clinical text. We evaluate the proposed model both intrinsically and extrinsically, and the results show that ClinicalT5 compares favorably with its close baselines. We also test upon more complicated patient outcomes prediction tasks, the results of which indicates its potential for these real-world downstream tasks in the clinical domain.

Limitations

In this work we present a generative language model for clinical texts based on T5. Although our experiments demonstrate the effectiveness of our method, there are still some limitations that can be improved in future work. First, our evaluation has not included question answering and other related tasks for clinical texts. These are important tasks (Phan et al., 2021) and can be further

explored in future work. Second, our pre-training method for ClinicalT5 has mainly inherited the objectives from T5 using direct unlabeled texts. As such, many important domain-specific knowledge for clinical domain (e.g., knowledge bases, concept definition) has not been explored to improve our generative model, serving as a promising direction for future research.

Ethics Statement

All datasets used in this research are publicly available and are obtained according to each dataset's respective data usage policy. We avoid showing any direct excerpts of the data in the paper. We do not attempt to identify or deanonymize users in the data in any way during our research, thus preventing any bias in our methods toward any specific users.

More specifically, the proposed models are trained on the clinical notes of the public MIMIC-III database, which are already deidentified in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting. As such, all identifying data elements in HIPAA, including patient name, telephone number, address, and dates, are already removed (Johnson et al., 2016) from our training data to hinder attempts to retrieve personal information from our models. Similar to existing pre-trained and publicly available models for the clinical domain, i.e., ClinicalBERT (Huang et al., 2019) and BioClinicalBERT (Alsentzer et al., 2019), the proposed models serve as a resource to facilitate future research on clinical text.

Acknowledgement

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112 and the NSF grant CNS-1747798 to the IUCRC Center for Big Learning. This research is also based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2019-19051600006 under the Better Extraction from Text Towards Enhanced Retrieval (BETTER) Program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ARO, ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Aishwarya Krishna Allada, Yuanxin Wang, Veni Jindal, Morteza Babee, Hamid R Tizhoosh, and Mark Crowley. 2021. Analysis of language embeddings for classification of unstructured pathology reports. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2378–2381. IEEE.

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 (ClinicalNLP)*, pages 72–78.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

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

Rajarshi Bhowmik, Karl Stratos, and Gerard de Melo. 2021. Fast and effective biomedical entity linking using a dual encoder. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 28–37, online. Association for Computational Linguistics.

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

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and

Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

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

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.

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

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

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

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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.

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.

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

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

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

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

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

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

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. In *AMIA annual symposium proceedings*, volume 2010, page 572. American Medical Informatics Association.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task (BioNLP)*, pages 58–65.

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

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.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers for open-domain question answering. 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 6648–6662, Online. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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

Xiao Zhang, Dejing Dou, and Ji Wu. 2020. Learning conceptual-contextual embeddings for medical text. In *AAAI*, pages 9579–9586.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Reproducibility Checklist

- **Source code with specification of all dependencies, including external libraries**: The models and source code along with a README file will be released upon decision of the paper.
- **Description of computing infrastructure used**: In this work, we use 3 Nvidia Tesla V100-32GB GPUs for the pre-training and one GPU for evaluations. PyTorch 1.8.1 and Huggingface-Transformer 4.18.0 (Wolf et al., 2019) are used for implementation.
- **Average runtime for each approach**: We pre-train the model on MIMIC-III for 3 epochs which takes $\sim$ 8 hours, and the best variant is chosen based on its performance on HOC.
- **Number of parameters in the model**: ClinicalT5-base has $\sim 220M$ parameters with 12 layers and ClinicalT5-large has $\sim 770M$ parameters with 24 layers.
- **Explanation of evaluation metrics used, with links to code**: We use the same measures and correctness criteria as in prior work (Zhang et al., 2019; Phan et al., 2021; Zhang et al., 2020; Lu et al., 2021c) for fair comparison. In particular, we use Pearson's and Spearman's correlation coefficients for intrinsic evaluation, and use precision, recall, F1 score as well as accuracy for extrinsic evaluation. We also use AUC of ROC, AUC of PRC and RP80 for the experiments of patient outcomes prediction.
- **Bounds for each hyper-parameter**: For all the experiments, we choose the learning rate from [1e-5, 1e-4, 1e-3] for the AdamW optimizer, the batch size from $[4, 8, 16]$.