PULSAR: Pre-training with Extracted Healthcare Terms for Summarising Patients' Problems and Data Augmentation with Black-box Large Language Models

Hao Li[♣*†], **Yuping Wu**[♣*], **Viktor Schlegel**[♢♣], **Riza Batista-Navarro**[♣]
Thanh-Tung Nguyen[♢] **Abhinav Ramesh Kashyap**[♢♡], **Xiaojun Zeng**[♣]
Daniel Beck[♠], **Stefan Winkler**[♢♡] and **Goran Nenadic**[♣]
[♣] University of Manchester, United Kingdom [♠]University of Melbourne, Australia
[♢]ASUS Intelligent Cloud Services (AICS), Singapore
[♡]National University of Singapore, Singapore

Abstract

Medical progress notes play a crucial role in documenting a patient's hospital journey, including his or her condition, treatment plan, and any updates for healthcare providers. Automatic summarisation of a patient's problems in the form of a "problem list" can aid stakeholders in understanding a patient's condition, reducing workload and cognitive bias. BioNLP 2023 Shared Task 1A focuses on generating a list of diagnoses and problems from the provider's progress notes during hospitalisation. In this paper, we introduce our proposed approach to this task, which integrates two complementary components [‡]. One component employs large language models (LLMs) for data augmentation; the other is an abstractive summarisation LLM with a novel pre-training objective for generating the patients' problems summarised as a list. Our approach was ranked second among all submissions to the shared task. The performance of our model on the development and test datasets shows that our approach is more robust on unknown data, with an improvement of up to 3.1 points over the same size of the larger model.

1 Introduction

Medical progress notes are used to document a patient's course in a hospital, including their current condition, treatment plan, and any updates to the plan (Li et al., 2022). Automated identification of treated problems from the assessment sections of a progress note in form of a "problem list" can help healthcare stakeholders to gain an accurate understanding of the patient's condition, reducing workload and cognitive bias (Gao et al., 2022a). This problem list is then used to outline and pursue a detailed treatment plan.

The majority of studies on clinical summarisation have focused on clinical notes; radiology reports (Zhang et al., 2018; MacAvaney et al., 2019; Gharebagh et al., 2020; Kondadadi et al., 2021; Dai et al., 2021), and progress notes (Moen et al., 2016; Liang et al., 2019; Adams et al., 2021; Gao et al., 2022a). In contrast, some studies have focused on dialogues (Yim and Yetisgen-Yildiz, 2021; Manas et al., 2021; Zhang et al., 2021). Recently, Gao et al. (2022b) proposed the task of "progress note understanding", where the goal is to generate problem lists given the assessment sections of a progress note. They further explored the performance of `T5` (Raffel et al., 2020), `BART` (Kondadadi et al., 2021) based on pre-training tasks with masked healthcare concepts (Gao et al., 2022a). To draw further attention to this task, The BioNLP 2023 Shared Task 1A (Gao et al., 2023) invited external participants to develop approaches to advance the state-of-the-art on the proposed task.

The main contribution of this work is a novel framework for data augmentation and summarisation of diagnoses/problems. In our approach, first, we optimise a domain-specific Language Model (LM) using a combination of different pre-training objectives, depicted in Figure 1; this model significantly outperforms the state-of-the-art, even when optimised on a limited number of manually annotated progress notes. Second, we instruct Large Language Models (LLMs) to generate synthetic data, in order to reduce the reliance on large, high-quality annotated datasets. Finally, we use the generated data to fine-tune the domain-specific LM on the task of problem list generation, given appropriate progress note sections. Our approach ranked second among all submissions to the shared task without additional annotated data. The results of our evaluation suggest that our pre-training objectives are aligned with the downstream task of summarisation and can significantly improve performance.

[*]These authors contributed equally to this work
[†]Corresponding email: `hao.li-2@manchester.ac.uk`
[‡]Our code is avaliable at https://github.com/yuping-wu/PULSAR

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 503–509

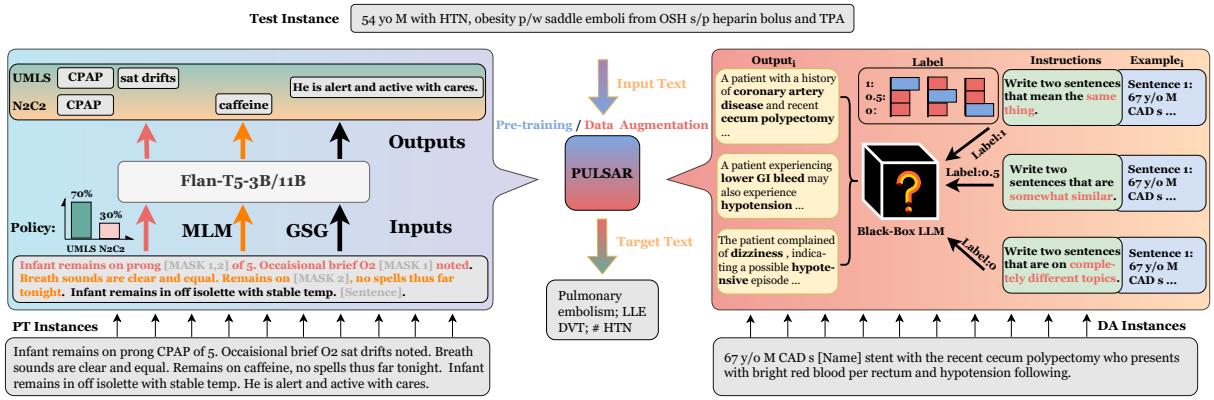

Figure 1: Overview of PULSAR. The left component represents the pre-training process with three different mask policies depicted in different colours. Both Gap Sentences Generation (GSG) and Masked Language Modelling (MLM) are applied simultaneously to this example as pre-training objectives. The right component shows the workflow for data augmentation where the three labels $\{1, 0.5, 0\}$ represent SAME THING, SOMEWHAT SIMILAR and COMPLETELY DIFFERENT TOPICS, respectively. PT INSTANCES and DA INSTANCES stand for PRE-TRAINING INSTANCES and DATA AUGMENTATION INSTANCES, respectively.

2 Methodology

Figure 1 shows the two components of our framework: first we pre-train an encoder-decoder model on MIMIC-III progress notes (Johnson et al., 2016) using three different concept masking pre-training objectives. Then we employ data augmentation when fine-tuning our model for the summarisation task.

2.1 Pre-training Model

The items on the problem list are not necessarily directly extracted from the original progress notes and hence we cast the problem as abstractive summarisation. Drawing inspiration from PEGASUS (Zhang et al., 2020a), we used an objective which closely resembles the abstractive summarisation objective, to gain better and faster fine-tuning performance.

Following the success obtained through masking words and contiguous spans (Joshi et al., 2020; Raffel et al., 2020), we propose to select and mask text spans or whole sentences from input documents. We concatenate these "gap text spans (sentences)" into a pseudo-summary. Gap text spans were selected by the QuickUMLS entity linking (Soldaini and Goharian, 2016) and an NER model trained on the i2b2-2010 challenge (Uzuner et al., 2011). Similar to the T5 pre-training procedure (Raffel et al., 2020), these text spans were replaced by "sentinel" mask tokens $< extra\_id\_i >$ to inform the model that input was masked. Here, i indicates the number of the mask (from left to right). The output sequence thus consists of the dropped-out text spans, delimited by the sentinel token between terms and the last $< extra\_id\_i >$ input representing the end of the output. Figure 2 illustrates our pre-training objective.

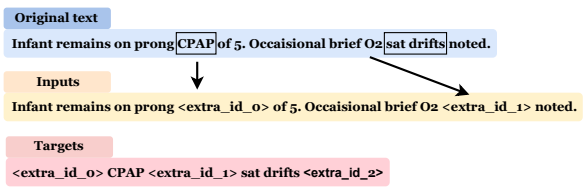

Figure 2: Our pre-training objective. The terms "CPAP" and "sat drifts" are identified by the NER models and replaced by a unique sentinel token respectively. The objective is to predict these masked-out spans.

Specifically, we considered three masking policies in our pre-training objective. For each sentence, When both tools identified entities, we selected UMLS terms with the probability of 0.7 and i2b2 terms with the probability of 0.3. When only one tool identified entities, these entities were selected. Finally, when no entities were identified, the entire sentence was masked with a probability of 0.15. In order to provide the model with the necessary medical knowledge and reduce domain barriers (Pandey et al., 2022), we leverage all progress notes from MIMIC-III (Johnson et al., 2016) to train `Flan-T5` (Chung et al., 2022) on this objective. The processed pre-training corpus had 2.08m rows of data, with 2.2k containing no UMLS terms, 23k containing no i2b2 entities, and 797 where neither tool recognised any entities.

Approach (Setting)	Dev set R-1/R-2/R-L	Test set R-F1/R-P/R-R
PULSAR$_{3B}(DA)$	**36.27/16.78/33.83**	30.48/38.02/29.72
PULSAR$_{11B}(DA)$	35.92/15.87/33.14	**31.15**/40.93/28.73
PULSAR$_{3B}$	33.60/13.70/31.32	31.14/**44.30**/27.18
PULSAR$_{11B}$	33.38/13.14/30.63	30.34/42.68/27.12
FlanT5$_{11B}(DA)$	32.57/13.07/29.95	-
FlanT5$_{11B}$	31.24/11.42/28.25	30.06/40.61/27.25
FlanT5$_{3B}(DA)$	29.46/09.85/26.15	30.47/38.01/29.72
ClinicalT5$_{\text{LARGE}}(DA)$	28.60/11.13/26.11	25.43/25.67/**32.05**
FlanT5$_{3B}$	28.90/08.93/25.26	30.60/41.09/28.58
PULSAR$_{3B}(-A)$	27.70/10.60/24.34	28.29/38.24/26.54
ClinicalT5$_{\text{LARGE}}$	31.09/12.85/28.15	19.92/18.93/28.89
FlanT5$_{\text{LARGE}}$	29.86/10.19/27.08	-

Table 1: Performance of evaluated models on the development set measured in terms of Rouge-1/2/LCS, and on the test set measured in terms of Rouge-F1/Precision/Recall, respectively. The composition of the input content is ASSESSMENT + SUBJECT + OBJECT, except where only the ASSESSMENT section of the input was used, indicated by -A. DA means that data augmentation was employed. The Rouge-L score on the development set was used for official ranking. Colours (i.e. 1st, 2nd, 3rd, 4th, 5th, 6th) indicate the highest to lowest performance.

Approach(MaxLen)	R-1/R-2/R-L
Baselines	
T5$_{\text{LARGE}}(512)$	29.901/10.81/28.21
FlanT5$_{\text{BASE}}(512)$	27.16/8.9435/24.90
ClinicalT5$_{\text{SCRATCH}}(512)$	26.68/9.51/23.94
T5$_{\text{BASE}}(512)$	25.07/7.72/23.36
FlanT5$_{\text{BASE}}(1024)$	25.51/7.96/23.07
ClinicalT5$_{\text{BASE}}(512)$	22.27/7.61/20.49
PEGASUS$_{\text{XSUM}}(512)$	22.39/6.86/20.36
ClinicalT5$_{\text{BASE}}(1024)$	21.13/7.19/19.55
ClinicalT5$_{\text{SCI}}(512)$	14.12/4.61/13.22

Table 2: Performance of baseline models on the development measured in terms of Rouge-1/2/LCS. The composition of the input content is ASSESSMENT + SUBJECT + OBJECT. The same colour represents the same model with different input lengths.

2.2 Data Augmentation (DA)

The lack of high-quality annotated data is a bottleneck that inhibits supervised learning methods in the healthcare field. For example, BioNLP Task 1A (Gao et al., 2023) has only 764 annotated training examples. Therefore, we rely on data augmentation techniques to obtain more training samples. Specifically, we propose a novel healthcare data generation (DG) framework based on DINO (Schick and Schütze, 2021; Li et al., 2023), which exploits the generative abilities of LLMs by relying on instruction following rather than model training. Our instructions to the LLMs include task-specific descriptions (i.e., "*Write two sentences that mean the same thing but keep these two healthcare terms* $[Term1], [Term2]$. Sentence 1: $[Source]$ Sentence 2: ") to make the model generate a paraphrase of $[Source]$, which is selected from the annotated training data. The instructions to keep terms aim to keep relevant terms from $[Source]$ which also appear in the problem list (i.e. the output). In addition, we might expect that the text generated by the LLM would only fit well into the corresponding instruction but would not be applicable as a reasonable output for other instructions. For example, in Figure 1 (i.e. label $\{1\}$ is the expected label in blue and label$\{0\}$ is the count label in red), it is expected that the generated text should have the same meaning as $[Source]$, but at the same time not have a completely different meaning from $[Source]$. Following previous work, we employ the self-debiasing (Schick et al., 2021) algorithm to achieve this objective, i.e. when predicting the next token, not only the probability of the corresponding label is considered, but also the counter label is taken into account. We then use BERTScore (Zhang et al., 2020b) and BLEURT (Sellam et al., 2020) to assess the similarity between each generated sample and the source, removing 85% of the lowest scoring generated sentences. The backbone of the framework can be any generative LLM, such as `GPT3.5`[§], `GPT3` (Brown et al., 2020) and `GPT2` (Radford et al., 2019) series models. Limited by the data use agreement, we used `BioMedLM` (Bolton et al., 2022), an open-source GPT-style model pre-trained on the biomedical abstracts and papers, [¶].

2.3 Implementation Details

Pre-training: We choose `FlanT5-3B` and `FlanT5-11B` (Chung et al., 2022) as our LM. `PULSAR-3B` and `PULSAR-11B` are pre-trained on two NVIDIA Tesla A100 80GB GPUs and four NVIDIA Tesla A100 80GB GPUs for 1 epoch respectively[¶]. During the pre-training, we rely on Fully Sharded Data Parallel (FSDP) with CPU offloading (Baines et al., 2021) to fit LLMs into GPU memory.

Data Augmentation: We employ `BioMedLM`

[§]chat.openai.com

[¶]The official test set result for PULSAR-11B was fine-tuned after the 0.33 pre-training epoch.

(Radford et al., 2019) as the data augmentation model with default settings, setting maximum output length to 40. Finally, the generated data are matched with the corresponding summaries, subjective and objective to create a training set of 1k instances. The DA model (Schick and Schütze, 2021) is run on a single NVIDIA Tesla V100 32G GPU, with each run taking up to twelve hours. Example templates and the full dataset description can be seen in Appendix A.

3 Experimental Setup

Baselines: We have chosen to adapt T5-base as one of our baselines, similar to the approach taken by Gao et al. (2022a). Additionally, we have incorporated various state-of-the-art models such as `FlanT5` (Chung et al., 2022), `ClinicalT5` (Lehman and Johnson, 2023) and `PEGASUS` (Zhang et al., 2020a). Whereas `FlanT5` is an enhanced version of T5 that has been finetuned in a mixture of tasks (Chung et al., 2022) and `ClinicalT5` pre-trained on MIMIC-III (Johnson et al., 2016). `PEGASUS` is an abstractive summarisation model with Gap Sentences Generation and Masked Language Model (Devlin et al., 2019) as pre-train tasks.

Evaluation metrics: We calculate ROUGE (Lin, 2004) scores on the test set, by comparing the generated summaries to their corresponding references, averaging for all generation samples. For all experiments, the data set was divided into a "train" and a "dev" set with a ratio of 8:2 for training and evaluation, respectively. The results are presented in Table 1, left column, and Table 2. Table 1, right column, shows the performance of the models on the official withheld test set. In this case, both train and dev splits were used for training.

4 Results and Analysis

Pre-training helps: Both Table 1 and Table 2 demonstrate that the pre-training objective improves task performance (compare 3B and 11B `PULSAR` to corresponding `FlanT5` models). The best performance of `PULSAR` was 3.1 points higher than the `FlanT5-11B` on the development set as the training set and 11.2 points higher than `ClinicalT5-large` on the official test set. The small difference in performance between `PULSAR-11B` and `PULSAR-3B` is primarily because the former has only completed 1/3 of the first pre-training epoch, potentially resulting in a lack of relevant medical knowledge and familiarity with downstream task patterns.

Data augmentation is effective when the data distribution is consistent; It is significantly more helpful for small models when on a random data distribution: Table 1 shows that, data augmentation improves performance (3 point on average, compared to not using DA). This shows that the proposed DA approach can effectively alleviate the lack of annotated healthcare data, when the distribution of training and test set data is consistent. From Table 1, it becomes evident that smaller models (`ClinicalT5-large`) can improve by up to 6 points with the help of data augmentation, but the effect diminishes with model size as it increases max to 2.5 on LLMs. The potential reason is that the test set for the sharing task differs significantly from the training set, in the vary of length of the summary.

The model is capable of discriminating irrelevant information, but longer input lengths may result in decreased performance: We conducted ablation experiments on `PULSAR-3B` to verify the impact of the input text type. In contrast to Gao et al. (2022b)'s findings on the small model, the results (`PULSAR-3B` vs. `PULSAR-3B-A`) in Table 1 show that if the input is ASSESSMENT + SUBJECTIVE + OBJECTIVE, the model performs better (by 2.9 points on the official test set and by 7 points on the development set) compared with only using ASSESSMENT as input. This indicates that while most of the relevant information can be inferred from the ASSESSMENT section alone, additional input can be beneficial. However, increasing the input length appears to not be useful: Table 2 shows that models trained with longer input lengths (1024 tokens) do not improve over models that were trained on 512-token-long input.

5 Conclusion

This paper contributed to the development of summarising patients' problems. Firstly, we proposed a novel task-specific pre-training LLM objective. Compared with other submissions, we rank 2nd at the official shared task without using additional manually annotated training samples. Secondly, we propose a new data augmentation framework and demonstrate its effectiveness in the healthcare domain. In the future, we will explore the applicability of our approach to other domain-specific generative tasks and conduct a deeper analysis of factors that contribute to overall model performance.

Limitations

The proposed model is computationally demanding. Recent work on parameter-efficient fine-tuning methods, such as LoRA (Hu et al., 2022), suggests that they can significantly reduce the number of trainable parameters at a minimal performance cost, which may help further democratise the development of domain- and task-specific models. In addition, as we continued to pretrain, to obtain the PULSAR models, their tokenizer was inherited from corresponding Flan-T5 model. Thus it does not contain domain-specific terminology, which may be a limitation in terms of representation density (i.e. frequent clinical terms may be split in multiple rare sub-tokens).

Ethics Statement

For the present work, we used an existing anonymised dataset from BioNLP 2023 Shared Task 1A without any data protection issues. In addition, data augmentation only uses an open-source, off-line model which is not offensive to the data user agreement that is shared with a third party.

Acknowledgements

We thank the anonymous reviewers from the BioNLP 2023 Shared Task for their valuable feedback. We would also like to acknowledge the use of the Computational Shared Facility at The University of Manchester.

References

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What's in a summary? laying the groundwork for advances in hospital-course summarization. In *NAACL-HLT*, pages 4794–4811. Association for Computational Linguistics.

Mandeep Baines, Shruti Bhosale, Vittorio Caggiano, Naman Goyal, Siddharth Goyal, Myle Ott, Benjamin Lefaudeux, Vitaliy Liptchinsky, Mike Rabbat, Sam Sheiffer, et al. 2021. Fairscale: A general purpose modular pytorch library for high performance and large scale training.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. 2022. Pubmedgpt 2.7b.

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, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey 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. In *NeurIPS*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Songtai Dai, Quan Wang, Yajuan Lyu, and Yong Zhu. 2021. BDKG at MEDIQA 2021: System report for the radiology report summarization task. In *BioNLP@NAACL-HLT*, pages 103–111. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew M. Churpek, and Majid Afshar. 2022a. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *COLING*, pages 2979–2991. International Committee on Computational Linguistics.

Yanjun Gao, Dmitriy Dligach, Timothy A. Miller, Samuel Tesch, Ryan Laffin, Matthew M. Churpek, and Majid Afshar. 2022b. Hierarchical annotation for building A suite of clinical natural language processing tasks: Progress note understanding. In *LREC*, pages 5484–5493. European Language Resources Association.

Yanjun Gao, Timothy Miller, Majid Afshar, and Dmitriy Dligach. 2023. Bionlp workshop 2023 shared task 1a: Problem list summarization. *"Proceedings of the 22nd Workshop on Biomedical Language Processing"*.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross W. Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *ACL*, pages 1899–1905. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

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

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

Ravi Kondadadi, Sahil Manchanda, Jason Ngo, and Ronan McCormack. 2021. Optum at MEDIQA 2021: Abstractive summarization of radiology reports using simple BART finetuning. In *BioNLP@NAACL-HLT*, pages 280–284. Association for Computational Linguistics.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text.

Hao Li, Viktor Schlegel, Riza Batista-Navarro, and Goran Nenadic. 2023. Do you hear the people sing? key point analysis via iterative clustering and abstractive summarisation. *arXiv preprint arXiv:2305.16000*.

Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlali, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, Richard Andrew Taylor, Harlan M. Krumholz, and Dragomir Radev. 2022. Neural natural language processing for unstructured data in electronic health records: A review. *Comput. Sci. Rev.*, 46:100511.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using ehr data. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 46–54.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish A. Talati, and Ross W. Filice. 2019. Ontology-aware clinical abstractive summarization. In *SIGIR*, pages 1013–1016. ACM.

Gaur Manas, Vamsi Aribandi, Ugur Kursuncu, Amanuel Alambo, Valerie L Shalin, Krishnaprasad Thirunarayan, Jonathan Beich, Meera Narasimhan, Amit Sheth, et al. 2021. Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study. *JMIR Mental Health*, 8(5):e20865.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artif. Intell. Medicine*, 67:25–37.

Babita Pandey, Devendra Kumar Pandey, Brijendra Pratap Mishra, and Wasiur Rhmann. 2022. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. *J. King Saud Univ. Comput. Inf. Sci.*, 34(8 Part A):5083–5099.

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

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *EMNLP (1)*, pages 6943–6951. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Trans. Assoc. Comput. Linguistics*, 9:1408–1424.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *ACL*, pages 7881–7892. Association for Computational Linguistics.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J. Am. Medical Informatics Assoc.*, 18(5):552–556.

Wen-wai Yim and Meliha Yetisgen-Yildiz. 2021. 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*, pages 10–20.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *EMNLP (Findings)*, pages 3693–3712. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *Louhi@EMNLP*, pages 204–213. Association for Computational Linguistics.

A Example Appendix

Example of data augmentation input and output

Instruction: "Task: Write two sentences that mean the same thing, but keep these two healthcare terms ['blood', 'hypotension']."

Sentence 1: "67 y/o M CAD s [Name] stent with recent cecum polypectomy who presents with bright red blood per rectum and hypotension following."

Sentence 2: "A patient with a history of coronary artery disease and recent cecum polypectomy is experiencing lower gastrointestinal bleeding and hypotension after undergoing preparation."

Instruction: "Task: Write two sentences that are somewhat similar, but keep these two healthcare terms ['blood', 'hypotension']."

Sentence 1: "67 y/o M CAD s [Name] stent with recent cecum polypectomy who presents with bright red blood per rectum and hypotension following."

Sentence 2: "A patient experiencing lower GI bleed may also experience hypotension due to the significant loss of blood from the lower gastrointestinal tract."

Instruction: "Task: Write two sentences that are on completely different topics, but keep these two healthcare terms ['blood', 'hypotension']."

Sentence 1: "67 y/o M CAD s [Name] stent with recent cecum polypectomy who presents with bright red blood per rectum and hypotension following."

Sentence 2: "The patient complained of dizziness and lightheadedness, indicating a possible hypotensive episode. On a different note, the recent research study showed a link between a high fiber diet and reduced lower blood pressure."

Figure 3: Continuation text generated by prompted learning data augmented methods with three different template descriptions. We chose to give input sentence 1 and generate only sentence 2, which helps to generate sentence similarity datasets.