Few-Shot Fine-Tuning SOTA Summarization Models for Medical Dialogues

David Fraile Navarro[1], **Mark Dras**[2], **Shlomo Berkovsky**[1]

[1]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
[2]Department of Computing, Macquarie University, Sydney, Australia
david.frailenavarro@hdr.mq.edu.au, {mark.dras, shlomo.berkovsky}@mq.edu.au

Abstract

Abstractive summarization of medical dialogues presents a challenge for standard training approaches given the paucity of suitable datasets. We explore the performance of state-of-the-art models with zero-shot and few-shot learning strategies and measure the impact of pre-training with general domain and dialogue specific text on the summarization performance.

1 Introduction

Clinical dialogues between patients and health professionals are among the core elements of the clinical encounter, containing most of the initial anamnesis questions, diagnostic information, treatment options, patient advice and counselling. Doctors usually summarize the content of these conversations into clinical notes, after each clinical visit or make use of expensive human medical scribes. As recent speech-recognition technologies show increasingly good performance (Chung et al., 2021; Zhang et al., 2020b), capturing these dialogues and generating abstractive summaries would help to reduce clinician load and improve patient care (Coiera et al., 2018).

Abstractive summarization has been one of the main challenges for NLP (Gupta and Gupta, 2019). The accuracy of abstractive summarization has improved over the past years due to the use of transformer-based, sequence to sequence (seq2seq) models (Aghajanyan et al., 2021; Raffel et al., 2019), larger training datasets and denser neural networks. Although several general-purpose datasets such as XSum (Narayan et al., 2018), CNN-DailyMail (Hermann et al., 2015), and SAMSUM (Gliwa et al., 2019) have been used for their training and development, few corpora exist that could be applied to the health scenario, medical terminology rich dialogues, with frequent interjections, ellipsis, and logical connections between semantic units (e.g., drug Y *treats* condition Z and not vice versa).

We fine-tuned several state-of-the-art (SOTA) models in a newly created medical dialogue dataset of 143 snippets, based on 27 general practice conversations paired with their respective summaries. We tested 10 transformer models to assess their performance in abstractive summarization of these dialogues. We learned that models pre-trained on general dialogues outperform baseline models. BART-based models were found to achieve the highest scores, although medical inconsistencies persisted in the generated summaries. In the future, we plan to perform further evaluations as the need for metrics that highlight inconsistencies in medical summaries remains unresolved.

2 Background

Training and fine-tuning NLP models for medical tasks has been a challenge, given the paucity of high-quality training data, although several initiatives such as MIMIC (Johnson et al., 2016) and n2c2 challenges (Henry et al., 2020) have advanced the field. Strategies to reduce dependence on large training datasets, such as transfer learning, have been explored (Fabbri et al., 2021a) to improve the model performance. Transformer-based models and their various implementations are well suited for transfer learning and fine-tuning with sparse datasets.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 254 - 266
July 10-15, 2022 ©2022 Association for Computational Linguistics

Additionally, zero-shot and few-shot approaches may help strike the balance between the model's performance, training time and training data requirements. Several recent developments have shown the effect of few-shot strategies in medical abstractive summarization (Goodwin et al., 2020) as well as in online medical dialogues (Nair et al., 2021).

Dialogue: Doctor: Okay. Thank you for seeing Jane Doe. Jane is a student here. She gives a history of intermittent ear pain, both ears, isn't it? Jane: Yeah, both ears. Doctor: Bilateral ear pains at night? Jane: Yep, and occasionally throughout the day. Doctor: Oh okay? Jane: Yeah. Not like the pain, just the pulsing. Doctor: Oh okay? Jane: Sorry, I mean. Doctor: For several years and also, mainly in your right ear, isn't it? Jane: The pulsing is in the right ear. The pain is in the left ear. Doctor: Oh, pain in your left? Jane: Sorry. I'm just thinking about it now. Doctor: Sorry. I thought. Doctor: It was both ears? Jane: I'm noticing, when I think about it, sorry, the pulsing is definitely more in the right ear. Doctor: Left ear pain and also right ear pulsing? Jane: But I don't know how else to describe it. Like that's. Doctor: No, no. We know exactly what you mean? Jane: It, yeah, like a. Doctor: A throbbing? Jane: It's like a, yeah, throbbing. Like a blood rush sort of. Doctor: Pulsation? Jane: Sensation. But not. Doctor: Okay. With throbbing? Jane: Obviously blood rush. Doctor: Throbbing in, for up to six months, maybe six months? Jane: Yeah. About, up to six months. Doctor: She looks very well, looks very well. Nil to find today, today. BP, what was it? I think it was 104. Doctor: Okay.

Summary: Jane has a history of bilateral ear pains at night in her left ear and pulsing, throbbing sensation in her right ear, like a blood rush, for

Box 1: Dialogue-summary example

Although few-shot and pre-training strategies have been studied separately, none have experimentally compared how these two interact in the medical dialogue domain and how different seq2seq models perform under these circumstances. In this work, we study how different few-shot strategies and pre-trained models affect the performance of abstractive summarization in medical dialogues.

3 Methods

3.1 Dataset

Our dataset consisted of 27 recorded conversations between general practitioners and patients collected by (Quiroz et al., 2020), where the data was used to characterize the structure and content of primary care consultations. These recordings took place at Primary Care facilities at Macquarie Health Clinics, Sydney, Australia. The conversations were professionally transcribed and anonymized. The conversations included in the dataset exceeded the token limit for existing language models (either 512 or 1024 tokens). Thus, we pre-processed the dataset by slicing the conversations into 400-word snippets. They were further processed to ensure that they contained semantically sound pairs of clinician-patient interactions, e.g., Doctor asks questions and Patient answers. A small number of snippets (less than 5%) were removed as they did not contain relevant medical information, such that the final dataset consisted of 143 snippets, containing 56,158 words. Box 1 shows a sample snippet. The dataset was partitioned using an 80-20 train-evaluation split. The training split was then subsequently split into further incremental few-shot sub-samples.

3.2 Annotation

A trained primary care physician with over 7 years of practical experience created summaries for all the snippets maintaining the following clinical information: medical information, medical advice, prescriptions, and general patient information. Annotation was performed by a single person; therefore, no inter-annotator agreement was calculated. Summaries varied in length between 17 and 158 words, as some snippets were more informative than others, with an average length of 68 words.

3.3 Models

Transformer-based models are currently the SOTA in several summarization benchmarks (Aghajanyan et al., 2020). We included the BART (Lewis et al., 2019), PEGASUS (Zhang et al., 2020a), and T5 (Raffel et al., 2019) families of models in our evaluation.

Among the various fine-tuned variants of these models, we included those having a fine-tuned version of the base model trained on the SAMSUM dialogue dataset (Gliwa et al., 2019). This dataset contains dialogues from various online chats, and it is one of the freely available dialogue summarization datasets. We harnessed the 'large' versions of these models. To explore medical transfer learning, we also included one model fine-tuned for PubMed summarization (Gupta et al., 2021). All the models included are available at HuggingFace[1]. Overall, 10 models were included in our evaluation: T5, T5$_{\text{SAMSUM}}$, BART, BART$_{\text{SAMSUM}}$, BART$_{\text{CNN-Dailymail}}$, BART$_{\text{CNN-SAMSUM}}$, PEGASUS, PEGASUS$_{\text{CNN-Dailymail}}$ PEGASUS$_{\text{CNN-Dailymail-SAMSUM}}$, and PEGASUS$_{\text{PUBMED}}$. The complete training strategy and best-fine-tuned models are available in our GitHub repository[2] and on the HuggingFace platform[3].

3.4 Fine-tuning strategy

We used the HuggingFace implementation of transformers and adapted their default fine-tuning scripts[4]. The default fine-tuning strategy consisted of training models for 3 epochs without further adjustments. Given the small size of the dataset, the evaluation split was only used at the end of the training and was not used to adjust the learning rate, which was set to the default value for each model. Initial analysis also showed that the loss value increased with additional training epochs. Therefore, to avoid overfitting, no further rounds of training were performed.

We implemented an incremental few-shot learning (FSL) strategy evaluating the models at zero-shot, and then incrementally fine-tuning pre-trained models with 10-shot, 20-shot, 50-shot, and the full dataset.

3.5 Metrics and evaluation

We quantitatively evaluated the summaries using the ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-L-sum) (Lin, 2004) for each model and FSL strategy. These were calculated immediately after training with the provided script in the 20% (29 snippets) that were held out for evaluation. We also computed the improvement over zero-shot learning (ZSL) for each model with each incremental FSL step.

For the qualitative evaluation, a small sample of 7 generated snippets was inspected by a clinician, aiming to analyse the semantic and medical accuracy of the generated summaries according to the following aspects: (1) assertion (e.g., information is correctly affirmed or negated); (2) major (e.g., symptom, diagnosis or treatment) or minor medical information missing; (3) medical coherence (e.g., wrong cause-and-effect relationship); and (4) contradicting advice (e.g., stop treatment instead of start treatment).

4 Results

4.1 Quantitative evaluation

All the models pre-trained with dialogues outperformed their base counterparts both in ZSL and across all the FSL steps, irrespective of the underlying model (T5, BART or PEGASUS). Table 1 shows the ZSL performance of the base models and dialogue *(SAMSUM)* pre-trained models. The best-performing model within each family is highlighted for each metric. Table 2 shows the performance of the models pre-trained with the full dataset of 114 snippets. Figure 1 shows ROUGE-1 score for all models being incrementally trained with 0, 10, 20, and 50 shots, and the full dataset.

Overall, BART-based models outperformed both T5 and PEGASUS, both for ZSL and 10, 20, and 50 FSL steps. Training on the full dataset, BART$_{\text{-CNN-SAMSUM}}$ scored highest for ROUGE-1 and ROUGE-2, but T5$_{\text{-SAMSUM}}$ outperformed it for the ROUGE-L and ROUGE-L-sum scores. Appendix A shows the full results across the FSL steps for all models.

Baseline	R-1	R-2	R-L	R-L-Sum
T5	30.93	11.40	22.44	28.59
T5$_{\text{-SAMSUM}}$	**35.74**	**13.99**	**24.63**	**33.76**
BART	32.70	9.69	19.74	30.78
BART$_{\text{-CNN}}$	36.72	11.90	22.46	34.73
BART$_{\text{-SAMSUM}}$	37.38	15.88	26.11	35.40
BART$_{\text{-CNN-SAMSUM}}$	**40.82**	**16.00**	**27.26**	**38.78**
PEGASUS	**35.23**	11.46	22.95	**32.83**
PEGASUS$_{\text{CNN}}$	34.36	12.06	23.66	29.68
PEGASUS$_{\text{CNN-SAMSUM}}$	33.69	**13.63**	**24.79**	31.79
PEGASUS$_{\text{-PUBMED}}$	15.31	1.00	10.41	13.99

Table 1: Zero-shot ROUGE scores

[1] https://huggingface.co/models

[2] https://github.com/dafraile/Clinical-Dialogue-Summarization

[3] https://huggingface.co/dafraile

[4] https://github.com/huggingface/transformers/tree/master/examples/pytorch/summarization

Full training (n=114)	**R-1**	**R-2**	**R-L**	**R-L-Sum**
T5	51.79	23.77	37.54	49.41
T5$_{\text{-SAMSUM}}$	**54.91**	**26.64**	**40.46**	**52.37**
BART	52.31	23.66	34.18	49.34
BART$_{\text{-CNN}}$	53.59	25.07	37.72	50.96
BART$_{\text{-SAMSUM}}$	52.99	24.88	37.22	50.87
BART$_{\text{-CNN-SAMSUM}}$	**55.32**	**27.12**	**39.67**	**52.22**
PEGASUS-	39.51	15.74	27.57	37.22
PEGASUS-$_{\text{-CNN}}$	**50.94**	23.30	36.40	48.52
PEGASUS-$_{\text{CNN-SAMSUM}}$	50.89	**24.54**	**37.25**	**48.92**
PEGASUS-$_{\text{-PUBMED}}$	30.87	11.13	21.05	28.30

Table 2 ROUGE scores for the full dataset training

Table 3 shows the average (across multiple models) relative improvement obtained for ZSL to FSL with 10, 20, and 50 shots, and the full dataset. This is further broken down into the baseline and dialogue-trained models. The performance of the models consistently improved with FSL increasing steadily up to 50-shot and further to the full dataset. The largest improvements were observed from baseline to 10-shot, and from 20-shot to 50-shot. Appendix B presents all the increments observed across FSL.

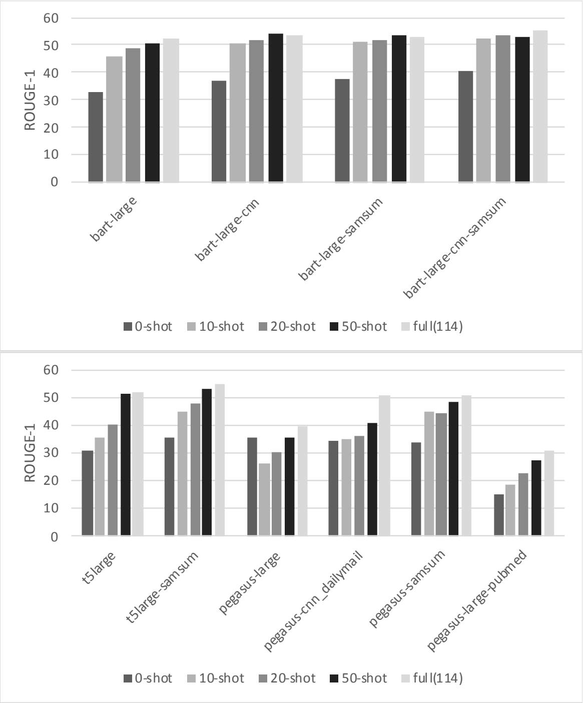

Figure 1: ROUGE-1 scores for each model for ZSL, FSL step, and the full dataset

Base Models	***ΔR-1***	***ΔR-2***	***ΔR-L***	***ΔR-Lsum***
10-shot	13.45	32.23	13.94	15.07
20-shot	22.70	49.71	23.60	26.13
50-shot	37.50	82.56	43.13	41.81
Full dataset	46.74	98.81	56.48	51.35
Dialogue Pre-Trained	***ΔR-1***	***ΔR-2***	***ΔR-L***	***ΔR-Lsum***
10-shot	30.69	52.52	35.32	31.43
20-shot	33.37	49.04	33.64	33.44
50-shot	41.43	67.31	46.20	43.38
Full dataset	45.49	74.17	50.65	46.83

Table 3: Average relative (%) improvement for the 4 ROUGE metrics and incremental FSL strategy

4.2 Qualitative evaluation

We focused our evaluation on the three best performing models of each family with respect to ROUGE scores: BART$_{\text{CNN-SAMSUM}}$, T5$_{\text{SAMSUM}}$, and PEGASUS$_{\text{-CNN-SAMSUM}}$. We detected several inconsistencies, incorrect advice, and missing information across the produced summaries. Box 2 shows a sample of the generated summaries, where the Doctor explores irritated tonsils caused by acid reflux and provides advice and treatment (Mylanta®).

PEGASUS$_{\text{-CNN-SAMSUM}}$
She is *coughing but not coughing*. Her tonsils are still swollen. If it is acid coming up, it can give her irritation in the throat. She hasn't lost a lot of weight, but she has lost one kilo. It is common to have acid problems related to spicy food, coffee, chocolate, alcohol and acidic food. (Missing: **no blood in bowel motions or vomiting, treatment advice, weight is missing**)

BART$_{\text{-CNN-SAMSUM}}$
No coughing, but the tonsils are still swollen. Acid can give you irritation in the throat. She hasn't lost a lot of weight. Fifty-four point three is one kilo less. There is no blood in bowel motions or vomiting blood. Sometimes it can be related to the food you eat spicy or acidic. You don't want to eat a huge amount of food in one sitting because it puts pressure on the stomach. **At night is pain at 11 when drinking hot milk.** ~~Tablets can also be bought from chemist Mylanta from Cambodia.~~

T5$_{\text{-SAMSUM}}$
Tonsils are still swollen and she is not coughing. Acid can give irritation in the throat. She hasn't lost a lot of weight. One kilo less. No blood in bowel motions or vomiting blood. Sometimes it can be related to spicy food, coffee, chocolate, alcohol or food that is very acidic. Trying to avoid certain foods. Drinking hot milk can sometimes help. There are tablets and also Mylanta liquid that she can drink. (Missing: 54.3 kg is the actual weight)

Human
No coughing but tonsils are still swollen. Acid coming up can irritate the lining of the throat. Weights 54.3 kilos and has lost a kilo. No blood in the bowel motions or when vomiting. Acid can relate to food you eat like spicy, coffee, chocolate, or alcohol. Sometimes a cup of hot milk helps. You can also buy Mylanta from the chemist.

Box 2: Sample of generated summaries and their evaluation. Legend: **Bold** – contradicting advice, *Italic* – medical incoherence, Underlined – missing information (*minor* or ***major***), ~~Strikethrough~~ – incorrect affirmation

Snippet evaluation (n=7)	T5-samsum	Pegasus-cnn-samsum	Bart-cnn-samsum
Missing information major	2	5	1
Missing information minor	2	2	1
Contradicting advice	0	0	1
Medical incoherencies	0	3	1
Assertion confused	1	3	0

Table 4: Qualitative examination of summaries

In the above examples, the best performing model, $\text{BART}_{\text{CNN-SAMSUM}}$ offered contradicting advice and incorrectly pointed out that the medicine needed to be bought in Cambodia (the country appeared in the text, but the meaning was confused). $\text{PEGASUS}_{\text{CNN-SAMSUM}}$ missed completely the medical advice given. $\text{T5}_{\text{SAMSUM}}$ did not produce incoherencies but failed to capture the actual patient weight. Table 4 shows the number of issues detected across the 7 examined snippets. Appendix C contains all the generated snippets and highlights additional issues.

5 Discussion

Our experiment shows that fine-tuning pre-trained models with few-shot learning offers a reliable strategy to improve summarization scores with small training data, making it appropriate for fine-tuning transformer models in domain-specific contexts, such as medical dialogues. By contrast, pre-training on medical literature did not improve results and showed poorer performance than the baseline models. BART based models achieve the highest ROUGE scores across all the FSL steps, with a relatively smaller footprint in terms of the required training time and the number of examples compared to both T5 and PEGASUS. Our experiment confirms previous findings that BART based models outperformed PEGASUS and T5 for summarization (Aghajanyan et al., 2020) and with few-shot strategies (Fabbri et al., 2021a). However, we observe that T5 gets higher ROUGE-L and ROUGE-L-sum results when trained on the full dataset. Although we obtain differences in the ROUGE scores across the best performing models, a limited qualitative analysis did not show a clear difference for T5 vs. BART. Our preliminary qualitative evaluation shows that T5 produced usable summaries (with no contradicting advice and no medical incoherencies) although further evaluation is required. This may reflect that relevant medical information may be situated at longer than 1-gram or 2-gram distances, suggesting that the longest common subsequence metric (ROUGE-L) can be more important for the quality of conversation summaries.

Moreover, we focus our analysis on the ROUGE score metrics, although this family of metrics alone is often insufficient to computationally appraise the quality of the summarization (Suleiman and Awajan, 2020). For instance, character n-gram F-score (chrF) (Popović, 2015), when evaluated for summarization tasks (Fabbri et al., 2021b) shows a higher correlation with the coherency of produced summaries than the ROUGE metrics. Further research is needed to establish the most apt metrics for evaluating the quality of medical summaries, especially as the need for maximizing factual correctness is critical for practical summarization applications in the medical domain.

An important limitation of our study is the small number of snippets and size of the medical dialogue dataset. Given the sensitive nature of medical conversations, this is a pervasive problem facing the development of NLP medical models. It is unlikely that medical dialogue conversations can be easily recorded, transcribed, and released as a public dataset given that they are likely to contain highly sensitive information. However, our experimental design focuses on this pervasive issue in medical NLP by exploring how FSL and pre-training may be leveraged to overcome the scarcity of large datasets.

In this work, we focus on a single document abstractive summarization. Given the length and complexity of medical dialogues, further experiments exploring multi-document summarization, aimed at producing full-dialogue summaries, would be necessary. Previous strategies for long-text summarization, such as global encoding seq2seq approach (Xi et al., 2020) or a globalized BERT architecture using a hierarchical propagation layer (Grail et al., 2021), may prove successful for summarizing long medical dialogues. Further model development, as well as refined training and fine-tuning strategies (e.g., adjusting transformer's structure, learning rate optimizations, and optimizing for additional metrics) or domain-specific dialogue datasets, may help further improve performance. Medical knowledge embeddings may also be a suitable strategy to improve performance and prevent medical incoherencies illustrated above.

Additional evaluations involving multiple clinicians and creating a more encompassing taxonomy of medical summarization errors would be needed for a thorough qualitative evaluation and proper appraisal of the model output quality. Establishing additional contrasts between qualitative and quantitative analysis may help to identify metrics that reliably capture important medical qualitative differences, potentially informing the development of new metrics, and quantifying the issues identified in our evaluation.

6 Conclusions and future work

Summarization of medical dialogues with FSL using pre-trained models is a feasible strategy for model development. Future research needs to focus on uncovering the most adequate set of metrics for capturing medically relevant and factually correct information in medical summaries. Additional qualitative evaluation may shed light on these issues and inform either the selection or development of the right metrics.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive Multi-task Representations with Pre-Finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training. *arXiv:2108.06209 [cs, eess]*, September. arXiv: 2108.06209.

Enrico Coiera, Baki Kocaballi, John Halamka, and Liliana Laranjo. 2018. The digital scribe. *NPJ digital medicine*, 1(1):1–5.

Alexander R. Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021a. Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation. *arXiv:2010.12836 [cs]*, April. arXiv: 2010.12836.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021b. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Travis R. Goodwin, Max E. Savery, and Dina Demner-Fushman. 2020. Flight of the PEGASUS? Comparing Transformers on Few-Shot and Zero-Shot Multi-document Abstractive Summarization. *Proceedings of COLING. International Conference on Computational Linguistics*, 2020:5640–5646, December.

Quentin Grail, Julien Perez, and Eric Gaussier. 2021. Globalizing BERT-based transformer architectures for long document summarization. In pages 1792–1810.

Som Gupta and S. K Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65, May.

Vivek Gupta, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. 2021. SUMPUBMED: Summarization Dataset of PubMed Scientific Articles. In pages 292–303.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Varun Nair, Namit Katariya, Xavier Amatriain, Ilya Valmianski, and Anitha Kannan. 2021. Adding more data does not always help: A study in medical conversation summarization with PEGASUS. *arXiv:2111.07564 [cs]*, November. arXiv: 2111.07564.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. *arXiv:1808.08745 [cs]*, August. arXiv: 1808.08745.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Juan C Quiroz, Liliana Laranjo, Ahmet Baki Kocaballi, Agustina Briatore, Shlomo Berkovsky, Dana Rezazadegan, and Enrico Coiera. 2020. Identifying relevant information in medical conversations to summarize a clinician-patient encounter. *Health Informatics Journal*, 26(4):2906–2914.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Dima Suleiman and Arafat Awajan. 2020. Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. *Mathematical problems in engineering*, 2020.

Xuefeng Xi, Zhou Pi, and Guodong Zhou. 2020. Global encoding for long chinese text summarization. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(6):1–17.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339. PMLR, November.

Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu. 2020b. Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition. *arXiv:2010.10504 [cs, eess]*, October. arXiv: 2010.10504.

1

Appendices

Appendix A: Results for 10, 20, and 50 few-shot strategies

Baseline(ZSL)	loss	rouge-1	rouge-2	rouge-L	rouge-Lsum
t5large	2.78	30.93	11.40	22.44	28.59
t5-large-samsum	2.26	35.74	13.99	24.63	33.76
bart-large	3.26	32.70	9.69	19.74	30.78
bart-large-cnn	2.15	36.72	11.90	22.46	34.73
bart-large-samsum	2.18	37.38	15.88	26.11	35.40
bart-large-cnn-samsum	2.00	40.82	16.00	27.26	38.78
pegasus-large	3.15	35.23	11.46	22.95	32.83
pegasus-large-cnn_dailymail	2.65	34.36	12.06	23.66	29.68
pegasus-large-cnn-samsum	2.20	33.69	13.63	24.79	31.79
pegasus-large-pubmed	6.93	15.31	1.00	10.41	13.99
10 shot	loss	rouge-1	rouge-2	rouge-L	rouge-Lsum
t5large	1.94	35.31	11.58	24.38	33.29
t5-large-samsum	1.79	44.81	20.05	33.55	42.73
bart-large	2.23	45.66	19.63	25.36	42.73
bart-large-cnn	1.95	50.79	23.22	34.90	48.17
bart-large-samsum	2.27	51.23	25.61	35.63	48.76
bart-large-cnn-samsum	1.97	52.28	26.18	37.84	49.65
pegasus-large	2.28	25.95	7.53	19.03	23.43
pegasus-large-cnn_dailymail	2.19	34.88	11.57	22.30	32.65
pegasus-large-cnn-samsum	1.93	44.56	19.33	32.19	42.41
pegasus-large-pubmed	4.99	18.81	2.74	13.22	17.20
20shot	loss	rouge-1	rouge-2	rouge-L	rouge-Lsum
t5large	1.68	40.37	15.46	29.05	38.17
t5-large-samsum	1.55	47.47	20.54	33.83	45.29
bart-large	2.20	48.89	22.05	27.17	47.01
bart-large-cnn	1.96	51.88	23.37	35.30	49.39
bart-large-samsum	2.22	51.65	24.21	34.79	49.11
bart-large-cnn-samsum	2.02	53.32	24.93	36.99	49.97
pegasus-large	2.09	30.49	9.82	20.94	28.52
pegasus-large-cnn_dailymail	2.05	36.31	12.44	24.23	34.22
pegasus-large-cnn-samsum	1.89	44.42	19.21	31.82	41.98
pegasus-large-pubmed	4.63	22.67	4.58	16.67	20.92
50shot	loss	rouge-1	rouge-2	rouge-L	rouge-Lsum
t5large	1.47	51.03	23.15	36.77	48.46
t5-large-samsum	1.43	53.19	25.02	38.98	51.07
bart-large	1.98	50.76	22.26	28.96	48.39
bart-large-cnn	2.11	54.06	26.80	39.18	51.79
bart-large-samsum	2.29	53.63	26.40	36.54	51.38
bart-large-cnn-samsum	2.10	53.15	25.19	38.99	51.11
pegasus-large	1.89	35.74	13.14	24.59	33.25
pegasus-large-cnn_dailymail	1.89	40.76	16.87	29.22	39.15
pegasus-large-cnn-samsum	1.81	48.27	22.71	35.60	46.20
pegasus-large-pubmed	4.07	27.29	8.69	18.56	24.85

all(114 shot)	loss	rouge-1	rouge-2	rouge-L	rouge-Lsum
t5large	1.39	51.79	23.77	37.54	49.41
t5-large-samsum	1.39	54.91	26.64	40.46	52.37
bart-large	1.86	52.31	23.66	34.18	49.34
bart-large-cnn	2.05	53.59	25.07	37.72	50.96
bart-large-samsum	2.05	52.99	24.88	37.22	50.87
bart-large-cnn-samsum	2.04	55.32	27.12	39.67	52.22
pegasus-large	1.78	39.51	15.74	27.57	37.22
pegasus-large-cnn_dailymail	1.81	50.94	23.30	36.40	48.52
pegasus-large-cnn-samsum	1.76	50.89	24.54	37.25	48.92
pegasus-large-pubmed	3.66	30.87	11.13	21.05	28.30

Appendix B: Relative (%) increase by training strategy for all models with 10, 20, 50 and full dataset

Model (10 shot % increase)	loss	rouge-1	rouge-2	rouge-L	rouge-Lsum
t5large	-30.08%	14.17%	1.60%	8.63%	16.44%
t5-large-samsum	-20.84%	25.38%	43.32%	36.19%	26.58%
bart-large	-31.42%	39.62%	102.71%	28.51%	38.84%
bart-large-cnn	-9.59%	38.31%	95.23%	55.42%	38.68%
bart-large-samsum	4.06%	37.05%	61.25%	36.47%	37.74%
bart-large-cnn-samsum	-1.62%	28.06%	63.65%	38.78%	28.03%
pegasus-large	-27.59%	-26.34%	-34.31%	-17.09%	-28.63%
pegasus-large-cnn_dailymail	-17.57%	1.52%	-4.06%	-5.76%	10.01%
pegasus-large-cnn-samsum	-12.25%	32.29%	41.87%	29.84%	33.39%
pegasus-large-pubmed	-27.97%	22.85%	173.45%	27.04%	22.89%
20 shot % increase	loss	rouge-1	rouge-2	rouge-L	rouge-Lsum
t5large	-39.51%	30.54%	35.64%	29.44%	33.50%
t5-large-samsum	-31.51%	32.81%	46.84%	37.31%	34.14%
bart-large	-32.34%	49.49%	127.63%	37.68%	52.73%
bart-large-cnn	-9.22%	41.27%	96.48%	57.21%	42.20%
bart-large-samsum	2.03%	38.18%	52.46%	33.26%	38.73%
bart-large-cnn-samsum	1.04%	30.61%	55.84%	35.66%	28.84%
pegasus-large	-33.62%	-13.45%	-14.33%	-8.76%	-13.11%
pegasus-large-cnn_dailymail	-22.60%	5.65%	3.15%	2.44%	15.31%
pegasus-large-cnn-samsum	-14.42%	31.88%	41.00%	28.32%	32.06%
pegasus-large-pubmed	-33.11%	48.09%	357.35%	60.19%	49.51%
50 shot % increase	loss	rouge-1	rouge-2	rouge-L	rouge-Lsum
t5large	-46.98%	65.00%	103.14%	63.84%	69.47%
t5-large-samsum	-36.83%	48.80%	78.88%	58.22%	51.28%
bart-large	-39.35%	55.20%	129.81%	46.72%	57.24%
bart-large-cnn	-2.16%	47.22%	125.27%	74.46%	49.09%
bart-large-samsum	5.15%	43.46%	66.26%	39.97%	45.15%
bart-large-cnn-samsum	5.20%	30.18%	57.43%	43.01%	31.79%
pegasus-large	-39.86%	1.46%	14.67%	7.12%	1.31%
pegasus-large-cnn_dailymail	-28.62%	18.62%	39.93%	23.51%	31.92%
pegasus-large-cnn-samsum	-17.79%	43.29%	66.69%	43.61%	45.30%
pegasus-large-pubmed	-41.32%	78.25%	767.70%	78.32%	77.56%
full dataset % increase	loss	rouge-1	rouge-2	rouge-L	rouge-Lsum
t5large	-49.78%	67.46%	108.51%	67.30%	72.81%
t5-large-samsum	-38.30%	53.61%	90.42%	64.25%	55.11%
bart-large	-42.90%	59.93%	144.22%	73.15%	60.32%
bart-large-cnn	-5.06%	45.92%	110.73%	67.97%	46.71%
bart-large-samsum	-5.96%	41.76%	56.65%	42.58%	43.70%
bart-large-cnn-samsum	1.79%	35.51%	69.54%	45.51%	34.64%
pegasus-large	-43.53%	12.16%	37.36%	20.12%	13.40%
pegasus-large-cnn_dailymail	-31.89%	48.24%	93.25%	53.88%	63.50%
pegasus-large-cnn-samsum	-20.00%	51.06%	80.05%	50.26%	53.86%
pegasus-large-pubmed	-47.18%	101.67%	1010.81%	102.24%	102.21%

Appendix C: Evaluation snippets and human summary.

model prediction_t5	model_prediction_pegasus	model_prediction_bart	human_summary
She was on Yas before and didn't get different headaches. She has not noticed a real difference in her headache since then. Sometimes hormones and migraines are related. Do you get worse when you are having your period or just before? Sometimes dehydration might trigger them. It is just the heat and if it rains it changes. A starting point would be to keep some sort of headache diary.	She didn't get more, different headaches on Yas or Estelle. It would have been nine months. Sometimes hormones and migraines are related. She does not get headaches worse when she is having a period or just before. On Sunday it was really hot and she was feeling basically dead.	She was on Yas before and didn't like it. She hasn't noticed any difference in headaches since then. Sometimes hormones and migraines are related. Not related to her period. Keep a headache calendar to see if there is a pattern. Try different treatments and also writing down the days and mark when periods are and exams are. Has no neck issues.	She was on Yas before, but she didn't liked it. She has been nine months on Estelle. She has not noticed a difference in headaches since then. She does not get worse headaches when having a period. She does not keep a calendar. On Sunday it was terrible, and it was really hot. A starting point is to keep a headache diary to see if there is a pattern. She does not have neck issues or pains.
She is in the boot still. She fractured one of the metatarsals and has avascular new necrosis on her talus. It took ages to figure it out because everyone thought it was the fracture that was painful.	She fractured one of the metatarsals and she has avascular new, new necrosis on her talus. She has been in the boot for 11 weeks and has an appointment with the orthopaedic surgeon in two weeks.	She fractured her toe and fractured one of the metatarsals. She has avascular new necrosis on her talus and has been in the boot for 11 weeks. Has an appointment with the orthopaedic surgeon in two weeks to check it out. Had a pelvic exam before. Is studying clinical science at Macquarie for two years. It is not covered by HECS.	Doing the swabs involves a pelvic exam and having a look at the external skin. She fractured her toe and is in the boot still. She fractured one of the metatarsals and have an avascular necrosis of the talus. She has been on the boot for 11 weeks. She has an appointment with the orthopaedic surgeon in two weeks. She studies clinical science at Macquarie. She does not know what she will do next.
He is coughing for two weeks. He feels itchy inside and there are sticky things in his throat. When he coughs it feels like a dry cough but it is not coming out. Breathing feels normal sometimes. Has asthma or chest problems. Used to use puffers. Had asthma four years ago and coughed all the winter. Have allergies to some food like yellow beans, beans and flour or some plant or seafood.	He has been coughing for two weeks. itchy inside of here and annoying sticky things in his throat. It feels like there is a bit of stuff there but it is not coming out. Has a history of asthma or chest problems. He used to have puffers but they are a long time ago. There was no fever or sore throat and nothing else. After he had ice cream he coughed a runny nose.	Cough has been two weeks. Apart from the cough he feels itchy. There are annoying sticky things in his throat. It feels like there is a bit of stuff there, but it is not coming out. Breathing feels normal sometimes. He had a history of asthma and chest problems in China and used to take puffers. Nothing else besides cough. No fever, sore throat, or runny nose. Cough reminds him of a cough a few years ago that didn't stop. Has allergies to some food but not heavy. Some kind of plant or seafood.	Cough for two weeks, feeling itchy inside and having stuff that is not coming out. Sometimes having difficulty breathing. Four years ago, had an asthma episode in China and used a puffer. No fever or sore throat. No runny nose. This cough reminds of the previous one. He says it start after a cold drink or ice-cream especially in winter. He has mild allergies to some food, usually pimples on the skin, but not affecting breathing.

Nothing like that. He is generally healthy. At night he coughs more but still has good sleep. Before this two weeks, if you are running or exercising, you get breathing problems or cough. Two weeks ago was in China and will start coughing after two ice creams. Temperature can sometimes have an effect on breathing. It could be that you picked up some sort of infection that has triggered the cough, but it is more like asthma. They will listen to his chest and use a puffer. There is no one around him.	He is generally healthy. He coughs more at night but has good sleep. If he is running or exercising he gets breathing problems or cough. Two weeks ago he was in China and had a cough one day before leaving. It doesn't sound like he has had the runny nose or sore throat.	He is generally healthy. He coughs more at night but still gets good sleep. Usually, he doesn't exercise or run. Two weeks ago he was in China and ate two ice creams. Temperature can have an effect on breathing but not like weeks of cough. It doesn't sound like he had a runny nose or sore throat. There are a lot of sticky things in the chest but it does not come out. They will have a listen to his chest and check his throat and ears and temperature. He may need to use a puffer for a while. Nobody else has a cough	He is healthy. Coughing more at night but still sleeping. He will cough when running for a train or bus or when exerting. He felt the cough was starting after eating an ice cream. Temperature can have an effect but won't give weeks of cough. He just has sticky things that doesn't come out. It could be some infection that triggered the cough. No one around him has cough.
It is difficult to breathe. She feels a little bit painful in the food pipe. Her digestion system is not functioning well. The intestine is fine. When she is hungry the acid level goes up and up. Last year she used to have that but she had some medicine to reduce the level in Cambodia. It helped when she took the medicine. Sometimes we do an endoscopy to look down with the camera. There is no vomiting or other stomach symptoms. No runny nose.	It feels like difficult to breathe. She feels pain in the food pipe. Her digestion system is not functioning well, so she thought it might be the stomach. The stomach is fine. Last year she used to have some medicine to reduce the acid level and it improved. It was prescribed in Cambodia.	It is difficult to breathe when breathing and a bit painful in the food pipe. The first time she met you said it was about the windpipe but because her digestion system is not functioning well, she thought it might be the front. There is nothing down here. She feels very hungry high up in her tummy and acid levels go up and up. Last year she used to have acid reflux but then she took some medicine to reduce acid level and it improved. It was prescribed in Cambodia. Taking the medicine helped. Sometimes we trial some treatment and if it responds it can be done with an endoscopy. Has not vomited or had a fever or runny nose.	It feels like is difficult to breathe and sometimes a little painful in the food pipe. Intestines and tummy feel fine. Feels uncomfortable when very hungry and that the acid is coming up. Used to have it last year but improved with medicine to reduce acid. It was prescribed in Cambodia. Sometimes we trial treatments, we don't rush in and do an endoscopy if not necessary. Felt tired. No vomiting, fever or runny nose.
Tonsils are still swollen and she is not coughing. Acid can give irritation in the throat. She hasn't lost a lot of weight. One kilo less. No blood in bowel motions or vomiting blood. Sometimes it can be related to spicy food, coffee, chocolate, alcohol or food that is very acidic. Trying to	She is coughing but not coughing. Her tonsils are still swollen. If it is acid coming up, it can give her irritation in the throat. She hasn't lost a lot of weight, but she has lost one kilo. It is common to have acid problems related to spicy food, coffee, chocolate, alcohol and acidic food.	No coughing, but the tonsils are still swollen. Acid can give you irritation in the throat. She hasn't lost a lot of weight. Fifty-four point three is one kilo less. There is no blood in bowel motions or vomiting blood. Sometimes it can be related to the food you eat spicy or acidic. You don't want to eat a huge amount of food in one sitting because it puts pressure on	No coughing but tonsils are still swollen. Acid coming up can irritate the lining of the throat. Weights 54.3 kilos and has lost a kilo. No blood in the bowel motions or when vomiting. Acid can relate to food you eat like spicy, coffee, chocolate, or alcohol. Sometimes a cup of hot

avoid certain foods. Drinking hot milk can sometimes help. There are tablets and also Mylanta liquid that she can drink.		the stomach. At night is pain at 11 when drinking hot milk. Tablets can also be bought from chemist Mylanta from Cambodia.	milk helps. You can also buy Mylanta from the chemist.
She had a miscarriage and gave some test on Tuesday. She started vaginal bleeding on last Wednesday and she was not here on Thursday. They gave her an ultrasound and they couldn't hear the heartbeat. The baby did not grow much.	She had a miscarriage. She was in study notes and couldn't do the test on time. Before last Wednesday she started bleeding, vaginal bleeding. On the Thursday she came to see the doctor and told them. They appointed her to Lucy and gave her ultrasound. The first ultrasound was 153 heartbeat. Then on the same day they could not find heartbeat and the baby didn't grow much.	She had a miscarriage. She was given a test on Tuesday and couldn't do it on time. Before last Wednesday she started bleeding, vaginal bleeding. They couldn't hear the heartbeat and she was about 13 weeks pregnant. The baby didn't grow much, about seven or eight weeks growth.	Lucy had a miscarriage. She started a vaginal bleeding last Wednesday. They did an emergency ultrasound and could not hear a heartbeat. First ultrasound was normal. She was 13 weeks pregnant. The first ultrasound showed it wasn't growing.

10