Leveraging Pretrained Models for Automatic Summarization of Doctor-Patient Conversations

Longxiang Zhang[1] **Renato Negrinho**[2] **Arindam Ghosh**[1] **Vasudevan Jagannathan**[1]
Hamid Reza Hassanzadeh[1] **Thomas Schaaf**[1] **Matthew R. Gormley**[2]

[1]3M Health Information Systems [2]Carnegie Mellon University
{lzhang28,aghosh4,juggy,hhassanzadeh,tschaaf}@mmm.com, {negrinho,mgormley}@cs.cmu.edu

Abstract

Fine-tuning pretrained models for automatically summarizing doctor-patient conversation transcripts presents many challenges: limited training data, significant domain shift, long and noisy transcripts, and high target summary variability. In this paper, we explore the feasibility of using pretrained transformer models for automatically summarizing doctor-patient conversations directly from transcripts. We show that fluent and adequate summaries can be generated with limited training data by fine-tuning BART on a specially constructed dataset. The resulting models greatly surpass the performance of an average human annotator and the quality of previous published work for the task. We evaluate multiple methods for handling long conversations, comparing them to the obvious baseline of truncating the conversation to fit the pretrained model length limit. We introduce a multistage approach that tackles the task by learning two fine-tuned models: one for summarizing conversation chunks into partial summaries, followed by one for rewriting the collection of partial summaries into a complete summary[1]. Using a carefully chosen fine-tuning dataset, this method is shown to be effective at handling longer conversations, improving the quality of generated summaries. We conduct both an automatic evaluation (through ROUGE and two concept-based metrics focusing on medical findings) and a human evaluation (through qualitative examples from literature, assessing hallucination, generalization, fluency, and general quality of the generated summaries).

1 Introduction

In recent years, pretrained transformer models (Lewis et al., 2019; Devlin et al., 2018; Zaheer et al., 2020; Brown et al., 2020) have been responsible for many breakthroughs in natural language processing (NLP) such as improved state-of-the-art performances for a broad range of tasks and the ability of training effective models for low-resource tasks. The demonstrated capability of transfer learning using large pretrained transformer models has led to widespread interest in leveraging these models in less standard NLP domains. Medical domains provide unique challenges and great potential for practical applications (e.g., see Amin-Nejad et al. (2020) and Huang et al. (2019)). Automatic generation of medical summaries from doctor-patient conversation transcripts presents several challenges such as the limited availability of supervised data, the substantial domain shift from the text typically used in pretraining, and potentially the long dialogues that exceed the length limitation of conventional transformers. Additionally, the model must have both extractive (e.g., such medications being taken, medication dosage, and numeric values of test results) and abstractive (e.g., the ability to determine the onset of a symptom from multiple conversation turns) capabilities.

Existing work on summarization from medical dialogue transcripts has achieved only limited success, both with pretrained models and otherwise. Krishna et al. (2020) relied on extra supervision to train a classifier to extract noteworthy utterances that are relevant to the target summary and do not handle the long conversations with their pretrained models, and their example results suffer from inferior fluency. Other existing work relying extractive methods is poorly adjusted to the informal nature of dialogue and the fact that information might not be present in any single span from the conversation transcript. Due to this, it has not yet been established that pretrained models are able to successfully perform automatic summarization from doctor-patient conversation transcripts.

In this paper, we attempt to tackle the task of medical dialogue summarization by leveraging pre-

[1]Code is available at https://github.com/negrinho/medical_conversation_summarization

trained transformer models. We show that BART (Lewis et al., 2019) can be fine-tuned to generate highly fluent summaries of surprisingly good quality even with a small dataset of no more than 1000 doctor-patient conversations (Section 2). We overcome the input length limitations through a multistage fine-tuning approach in which the task of dialogue summarization is achieved in two steps: summarizing portions of input conversation and rewriting aggregated summaries of each portion (Section 3). Our approach is simple as it amounts to fine-tuning pretrained model on appropriately constructed datasets. Despite its simplicity, it is effective at improving performance according to both automatic evaluation and human inspection (Section 4.1-4.3) when compared to the baseline approach of simply truncating the input. We also observe good generalization of our fine-tuned models across medical domains and conversation lengths, as shown by example conversations from other papers tackling the same task such as Krishna et al. (2020) and Joshi et al. (2020) (Section A.4). These examples also show the superior quality of our generated summaries.

2 Dataset

The dataset used in this paper is based on a collection of more than 80000 de-identified doctor-patient conversations (both audio and transcript). 1342 conversations of two major specialties: internal medicine and primary care are annotated by medical scribes using our annotation environment specifically designed for the task. The scribes listen to the conversation audio and fill in necessary information in a simulated Electronic Health Record (EHR) system. The EHR simulator consists of 14 distinct sections such as History of Present Illness (HPI) and Review of System (ROS).

We collect multiple references for each conversation, for a total of 21588 annotations. The dataset is split by conversation into train, development, and test with 939(15043), 201(3095), and 202(3450), respectively, where the values in parentheses are the number of HPI summaries in that split. Additional statistics are included in Appendix A.1.

We choose to use only the HPI section as our training target due to several observations: first, non-HPI sections are much less frequently filled by scribes, e.g., no more than 5% of all annotations have covered ROS section; second, scribes are required to write coherent paragraphs in the HPI section, whereas other sections might be structured as forms with most items being multiple choice; third, scribes are trained to cover non-HPI aspects like medication or physical examination in the HPI section if they are relevant to the "present illness" of the patient, making HPI section a good candidate for capturing most important medical findings in the conversation.

Each conversation in our dataset has on average 15 reference HPI summaries from different scribes. One running example of a long conversation (with more than 2200 words) and three corresponding references are showcased in Appendix A.3. As can be seen from the example, different references can exhibit large variance in length and quality. For consistency, we select target reference summaries in the training set as follows: first, we leverage our rule-based system to extract medical findings from all reference summaries; then we select the reference with the most findings as target. While filtering out low-quality training summaries is expected to impact the performance of the fine-tuned models, we leave such study for future work.

3 Methods

The methods that follow can be broken down into single-stage and multistage. All models rely on fine-tuning of pretrained BART models, the difference being how the datasets used for fine-tuning are constructed. For the single-stage approach, conversation transcripts are serialized with doctor and patient roles annotated (i.e., the encoder consumes a single sequence for the conversation) and mapped directly to the target summary. Conversations longer than the transformer model length limit are simply truncated, leading to unrecoverable information loss. Despite the simplicity of this approach, it works remarkably well and it serves as a strong baseline to beat. For the multistage approach, the conversation is first broken down into parts that are summarized independently by one model, and the resulting partial summaries are then aggregated and summarized into a final summary by another model. The methods that we propose in this class differ in how they break down the conversation into parts and therefore, the datasets that are used for fine-tuning their first stage model.

The multistage approach is motivated by the necessity of getting around the limited length budget of pretrained models along with the belief that medical findings covered in a summary is often present

locally in a contiguous set of turns between the doctor and the patient, allowing each part to be summarized independently, with a later aggregation stage of all part summaries.

3.1 Multistage summarization

We experiment with two methods of breaking down the conversations into parts and setting up datasets for fine-tuning the first stage summarizer:

SentBERT. We break all reference HPI summaries into individual sentences using the standard sentence splitter from the NLTK library (Bird and Klein, 2009) and then create a collection of snippets of eight consecutive turns by sliding window over the conversation with stride one. Cosine similarity between each summary sentence and all the snippets is then calculated using their respective hidden representations generated by the pretrained Sentence-BERT model (Reimers and Gurevych, 2019). All snippets that have a similarity of 0.7[2] or higher are then coalesced in case of overlap, and the longest such snippet is matched to the summary sentence. 99.6% of snippets generated in this way are within the input length limit, with an average of 230 tokens. One disadvantage of this method is that at inference time, we do not have reference summary to identify "similar" snippets. Therefore, the input to the second summarizer is created by first breaking each conversation into a set of 8-turn snippets with four turn overlap, and then generating single-sentence summaries from these snippets, and finally concatenating all generated sentences into a single paragraph[3]. See Figure 1 for an illustration on the inference procedure and Figure A.6 for examples of sentences generated for snippets.

Chunking. We create chunks of transcript from each conversation where each chunk consists of two components: a fixed-length "header" that is selected from the beginning of the conversation, and is present in all chunks; a variable "body" that is created by a sliding scan of the rest of the conversation. A special ellipsis token "..." is added between any header and body that are not contiguous and at the end of every non-terminating chunk, marking the existence of transcript text that is not present in the chunk. Each chunk is created to not exceed 512 words (approximately 800 tokens). The length of the chunk in number of words was chosen such that running the tokenizer of the pretrained model will result in a sequence that fits within its 1024 token length limit. The header length is chosen to be 128 words (c.f. hyperparameter tuning on the length of header in Appendix A.5), representing approximately 25% of the chunk. The target for every chunk from the same conversation is the complete HPI summary. Contrary to SentBERT, no special care is taken for constructing the summary targets for the chunks (i.e., we use the final desired summary for the conversation) as it is hypothesized the model will learn to only generate information if it is present in the chunk. See Figure 2 for an illustration and Figure A.5 for example summaries generated from conversation chunks.

Our simple multistage approach is proven to be effective in dealing with long conversations. As can be seen in Figure 3, 65.3% of the 939 conversations in the training set exceed the 1024 token limit and 35.5% exceed 2048 tokens; whereas only less than 10% of the inputs in the second stage of multistage fine-tuning have to be truncated, regardless of which method we use in the first stage. As we show in Section 4.3, overcoming the truncation problem can help generate summaries that cover information that occurs later in the conversation and have reduced hallucination.

3.2 Training

We leverage the pretrained BART model (Lewis et al., 2019) as our main model for summarization and we choose the model checkpoint pretrained on a BART large model (12 encoder and decoder layers, BART-LARGE, 405 million parameters) as the starting point for all our fine-tuning experiments. For comparison, we also use BigBird (Zaheer et al., 2020) with two different model checkpoints: one pretrained using RoBERTa (ROBERTA-BASE, 155 million parameters) (Liu et al., 2019) and one from Pegasus (PEGASUS-LARGE, 575 million parameters) (Zhang et al., 2020). Token limit for all models is set at 1024.[4]

The BART experiments are run using

[2] 0.7 as the similarity threshold leads to most reasonable snippets by sample inspection.

[3] No additional post-processing steps are taken to filter out "noisy" sentences from potentially irrelevant snippets, we hypothesize that training in the second stage should instruct the summarizer on how to filter out those sentences automatically

[4] On Nvidia Titan X Pascal GPU with 12GB memory, we experienced out-of-memory error when using BigBird with a token limit of 2048 or higher, therefore we decided to stay with the default token limit of 1024 and full attention calculation; this means the RoBERTa and Pegasus model checkpoints effectively reduce the BigBird model to ROBERTA-BASE and PEGASUS-LARGE models, respectively.

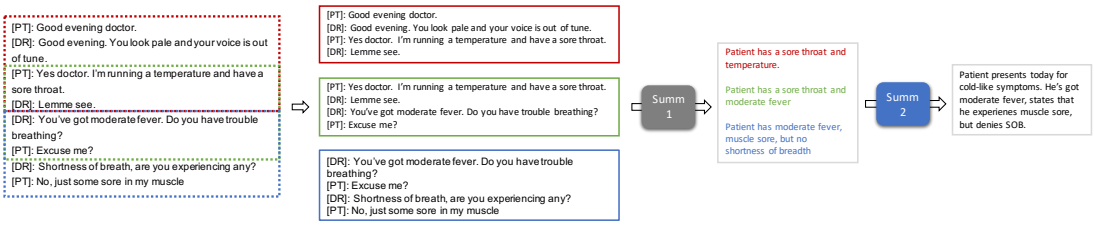

Figure 1: Multistage inference with SentBERT method. **Summ** stands for summarizer. The training target for **Summ 1** is a single sentence from the HPI summary. Complete summaries are used as target for **Summ 2** only.

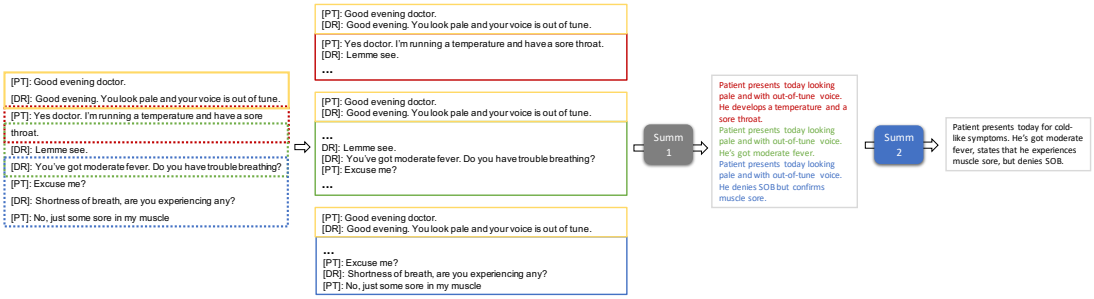

Figure 2: Multistage inference with Chunking method. **Summ** stands for summarizer. The same header (denoted by the yellow box) is added to the beginning of every chunk, serving as context, and the complete summaries are used as targets for fine-tuning both **Summ 1** and **Summ 2**.

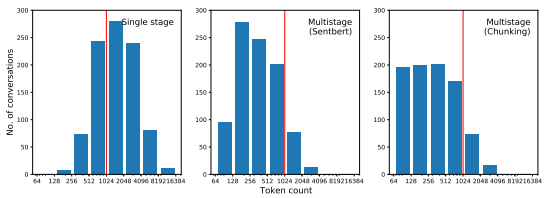

Figure 3: Token count histogram for original conversation (left), input to the second stage of multistage fine-tuning from using SentBERT (middle) and Chunking (right) datasets in the first stage. Vertical lines represent the 1024 token limit after which truncation occurs.

fairseq (Ott et al., 2019), while the BigBird experiments are run using the code released by the authors [5]. For all fine-tuning experiments, we follow the recommended procedures outlined in their respective repos. We choose the default BPE tokenizer for tokenization with a vocabulary size of 50264. Newline and tab characters in each conversation are replaced by whitespace and no further preprocessing is done. More hyperparameters are shown in Table A.1. The same hyperparameters are used in both single-stage and multistage fine-tuning. Model checkpoints are saved per epoch. After training, we run model inference on a subset of the development set to pick the checkpoint with the best ROUGE-1 F1 score as the candidate for further evaluation. For single stage fine-tuning on 939 conversations, training is usually finished within 10 epochs.

4 Experiments

We adopt ROUGE (Lin, 2004) as our main evaluation metric. Although ROUGE score has limited capability of capturing semantic similarities such as paraphrasing, which is common in abstractive summarization, we still consider it a useful metric for medical summarization due to restricted and highly technical vocabulary used in the medical domain. All references in dev and test set are used in automatic evaluation.

To address the limitation of ROUGE, we also introduce an automatic concept-based evaluation metric: medically relevant findings are first extracted from both generated and reference summaries by an external NLP system, and then precision, recall, and F1 score are calculated between the two sets of findings. Medical concepts are extracted via one of two systems: our in-house rule-based system and quickUMLS (Soldaini and Goharian, 2016). quickUMLS is a Python implementation of Unified Medical Language System (UMLS)[6] that standardizes various health and biomedical vocabularies. It is publicly available, and is capable of extracting a

[5] https://github.com/google-research/bigbird

[6] https://www.nlm.nih.gov/research/umls/index.html

	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
BART (large model, single stage)	**0.3029 (0.4364)**	**0.1047 (0.1841)**	**0.3191 (0.4285)**
BigBird (ROBERTA-BASE)	0.1697 (0.3297)	0.0633 (0.1662)	0.1933 (0.3600)
BigBird (PEGASUS-LARGE)	0.2570 (0.3949)	0.0822 (0.1889)	0.2669 (0.3964)

Table 1: ROUGE evaluation across models on dev set. Numbers in parentheses are "mean-of-best" ROUGE scores. Overall, the results obtained with BigBird were much worse than those obtained with BART, showing the importance of picking an appropriate pretrained model for fine-tuning.

wide scope of medical findings such as symptoms, diseases, medication and procedures. Our rule-based system is a commercial system proven to be effective at capturing symptom-related findings in clinical reports. Example concepts extracted from reference and generated summaries are shown in Appendix A.3, Figure A.4. False positive error is a major limitation of using those NLP systems, and is more severe with quickUMLS. We therefore implement majority voting to filter medical findings to be included in the reference set: any finding is included only if it is present in at least three human written summaries (or all of them when there are fewer references). The concept-based evaluation based on filtered findings is still susceptible to false positive errors, nevertheless, it provides an alternative to ROUGE as a potentially direct measure of the medical information coverage in generated summaries. Such a measure aligns better with the end-user (i.e., doctors) expectation of the summary quality. We leave research on better metrics for medical summarization as future work.

For automatic evaluation, we present results on the test set. Results on the development set can be found in Appendix A.5. As a more direct approach to quality assessment, we also conduct manual evaluation on a small sample of 10 conversations in the development set.

4.1 Pretrained model comparison

ROUGE scores for generated summaries across three models: BART, BigBird (RoBERTa), BigBird (Pegasus), are presented in Table 1. A "mean-of-mean" ROUGE score is calculated by first averaging the scores between the generated summary and all reference summaries for one conversation, and then averaging across conversations. Considering the variance in the length and quality of multiple references, we also calculate a "mean-of-best" ROUGE score: for each conversation, we pick the reference that scores the highest ROUGE-1 F1 with the generated summary and calculate other types of ROUGE scores; we then average the scores across conversations. BART strongly outperformed BigBird with either Roberta or Pegasus checkpoints. Upon manual inspection, we discovered that summaries generated with the BigBird models, or effectively ROBERTA-BASE and PEGASUS-LARGE, lack fluency and contain large amounts of repetition with sentences such as *The patient is here for a follow up follow up follow up ...*[7]. We choose to focus on BART in the remaining of the paper.

4.2 Automatic evaluation

ROUGE scores for both single-stage and multistage fine-tuning are shown in Table 2. Table 3 shows results for the concept-based evaluation. The Multistage (Chunking) method performs the best by ROUGE metrics, whereas concept-based evaluation leads to mixed results. Differences between the two concept-based evaluations are to be expected considering the different medical findings they cover. It is also worth noting that neither metric moves in unison with ROUGE, we therefore choose to view the three metrics as complementary and providing a more comprehensive interpretation of the quality of the generated summaries.

Multistage (SentBERT) method does not consistently improve on single-stage training, which could be attributed in part to the mismatch between snippets used in fine-tuning the first stage model and the snippets used for generating single sentence summaries as inputs for the second stage. For example, in Figure A.6 of Appendix A.5, we see that some snippets do not contain any noteworthy medical information. The number of such snippets is much larger for the SentBERT method than the Chunking method because of the small span of each snippet and the small stride used to

[7] We believe this is not due to different target length settings during inference. We have experimented with 128 and 256 target length for BART as well, and the drop in ROUGE score with shorter target length is no more than 10%. Model capacity may not explain the difference in performance either, as BigBird (Pegasus) model contains 40% more trainable parameters than that of BART.

slide over the conversation. This can lead to much noisier inputs for the second stage fine-tuning with more summarizing sentences potentially hallucinating medical contents, that are then unable to be effectively denoised by the second stage model.

In the last two rows in Table 2, we show two baseline evaluations to place other ROUGE scores in context. **training** computes ROUGE between generated summaries and a set of random target summaries in the training set. The approximate 20% drop in performance provides evidence that the model is not simply memorizing sentences from the training set. This is an important concern with medical summarization considering the intrinsic similarity between summaries of the same medical specialty (e.g., similarity among patients with diabetes). **reference** computes the average ROUGE scores measured among reference summaries. Specifically, for any conversation with multiple references, we do the same ROUGE evaluation used in the rest of the paper by treating each reference in turn as the generated summary and the remaining ones as targets. **reference** shows the worst scores of all experiments. Although this does not guarantee that the generated summaries by the model exceed human performance, we show through the running example and in Section 4.3 that model generated summaries can consistently be better than some reference human summaries.

Figure 4 shows the performance breakdown of all three methods by number of input tokens. We group all conversations in the test set into five buckets by their number of input tokens and compare for all methods both the "mean-of-mean" ROUGE scores (top row) and concept-based F1/P/R (bottom row) using quickUMLS. Multistage (Chunking) method outperforms the single stage model consistently across all buckets, even on conversations with fewer than 512 tokens, i.e. conversations that induce only one chunk in the multistage processing; however, the largest improvement in ROUGE score occurs for conversations in the (512, 1024] bucket, which is still within the input token limit of BART model, and we observe similar degradation in ROUGE scores across all three methods as conversation becomes longer. Concept-based evaluation, however, paints a different picture where improvement over single stage method is more significant for conversations beyond the 1024 token limit, which can be largely attributed to improved recall of concepts (see the third and fourth buckets, bottom center plot, Figure 4). Multistage (SentBERT) method also shows large improvement in concept-based evaluation for very long conversations (larger than 2048 tokens). This suggests both multistage methods lead to more reference medical concepts being generated, which may be favored over minor improvement in ROUGE score in the domain of medical summarization. Multistage (Chunking) method displays the most consistent improvement on conversations in the (1024, 2048] bucket across all types of evaluation metrics, one explanation could be that although multistage training can help circumvent information loss due to truncation, the input to the second stage, namely the concatenation of first stage summaries from all chunks, is also noisy; second stage performance on rewriting such a noisy input could degrade if the level of noise, or the number of first stage summaries, is too large.

4.3 Human evaluation

We employ two domain experts to conduct quality evaluation on 10 conversations in the development set. Five short conversations (less than 1024 tokens) and five long conversations (greater than 2048 tokens) are randomly chosen. For each conversation, we include summaries generated from both single-stage and multistage fine-tuning, as well as three reference summaries. One of the three references is selected as the one containing the most symptoms as extracted by our rule-based system (**reference (max. symp.)** in Table 4). The following factors are considered during evaluation:

- *fluency*: How fluent is the text generated?
- *relevancy*: Are contents relevant for HPI?
- *missing*: Are any key findings missing?
- *hallucination:* Are any findings hallucinated or inaccurate?
- *repetition*: Are there repetitive sentences?
- *contradiction*: Are any sentences contradicting each other?

Gender mismatch is not considered in the human evaluation as it was observed that, while the model frequently infers the wrong gender pronouns due to the lack of gender information in the transcript, it is sufficient to prefix the conversation with a sentence describing the desired gender for generations to use the correct pronouns. This would allow the development of a system that conditions on self-identified gender information for generation. See Appendix A.5 for an exploratory experiment.

	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
single stage	0.3131 (0.4427)	0.1097 (0.1819)	0.3281 (0.4337)
multistage (Chunking)	**0.3331 (0.4674)**	**0.1188 (0.1958)**	**0.3412 (0.4486)**
multistage (SentBERT)	0.3073 (0.4406)	0.1043 (0.1772)	0.3170 (0.4218)
training	0.2445 (0.3628)	0.0588 (0.1198)	0.2347 (0.3159)
reference	0.2920 (0.4239)	0.0852 (0.1638)	0.2932 (0.4083)

Table 2: ROUGE evaluation for BART fine-tuning on the test set. Values in parentheses are "mean-of-best" scores.

quickUMLS	F1	Precision	Recall
single stage	**0.4093**	0.5212	0.4009
multistage (Chunking)	0.4052	**0.5316**	0.3948
multistage (SentBERT)	0.4001	0.4813	**0.4166**
rule-based	F1	Precision	Recall
single stage	0.3617	**0.6410**	0.4112
multistage (Chunking)	**0.3847**	0.5951	0.4387
multistage (SentBERT)	0.3673	0.5135	**0.4622**

Table 3: Concept-based evaluation on test set.

Inter-rater agreement We calculate the Pearson's correlation coefficient ($\rho = 0.63$), Kendall rank correlation coefficient ($\tau = 0.51$) and Cohen's kappa ($\kappa = 0.22$) between the two domain experts as measures for inter-rater agreement. The low kappa score should be taken with a grain of salt because of frequent ties in the scores and tie breaking is done somewhat arbitrary during kappa calculation, we therefore focus more on the other two correlation coefficients and consider the agreement between the experts reasonable, but it does reflect the challenge in the consistency of quality evaluation for medical summaries even for experts.

Qualitative findings Table 4 shows the human evaluation scores for all summaries. Scores from the experts are averaged by experts and conversations. **reference (other)** stands for average score assigned to the other two references in each conversation. The difference in quality across generated and reference summaries are minor in *fluency*, *repetition* and *contradiction*, which indicates the generated summaries are as readable as those written by a human scribe. Generated summaries tend to score lower than the best human reference in *missing* and *hallucination*, with *missing* score being the lowest among all quality factors, suggesting that the fine-tuned models incur more frequently false negative errors. Surprisingly, scores of generated summaries are higher than **reference (other)** in *relevancy* and *missing* factors. This may be due to the large variability in quality across human references, but does provide encouraging evidence on the potential of using pretrained transformer models towards practical medical dialogue summarization.

Single stage fine-tuning leads to summaries with relevancy comparable to summaries generated by multistage fine-tuning, but with much worse hallucination score. At least among these 10 examples, we do not observe a clear difference in quality between summaries generated by both multistage methods. Hallucination in the single-stage model is more prevalent in longer conversations. For example, in Figure A.3 in Appendix A.3, the latter half of the single-stage summary starting from *She has a history of hyperlipidemia...* is largely an hallucination. We believe that this is partly due to the loss of information incurred by truncation (the example conversation contains around 2200 words, or approximately 3500 tokens), resulting in a model that learns to fill in frequently co-occurring information, even if it is not available in the truncated conversation transcript. Multistage summaries, on the other hand, successfully capture contents beyond the 1024 token limit in the conversation, such as medication like *Cialis*. It is also encouraging to see that the large amount of chitchat (see, for example, the last chunk in Figure A.5) present in the conversation is largely ignored in the generated summaries from multistage fine-tuning.

Generalization As a qualitative comparison with similar work in the field of medical dialogue summarization, we run inference with our fine-tuned models on conversations copied from Krishna et al. (2020) and Joshi et al. (2020). The results are shown in Appendix A.4. We include only summaries generated by our single-stage model as all example conversations are well within the 1024

	fluency	*relevancy*	*missing*	*hallucination*	*repetition*	*contradiction*
single stage	5.0000	4.7625	3.7375	3.8750	5.0000	4.6875
multistage (Chunking)	4.9375	4.6000	3.6875	4.2125	4.9250	4.7250
multistage (SentBERT)	4.9375	4.5375	4.0000	4.2000	4.8625	4.7500
reference (other)	4.8438	4.5313	2.7813	4.9313	5.0000	5.0000
reference (max. symp.)	4.9375	5.0000	4.6125	4.7250	5.0000	5.0000

Table 4: Human evaluation scores on ten conversations. Evaluated on a 5-point scale (higher is better).

token limit. Summaries generated by the multistage models are of comparable quality. The reference summary (Figure A.7) from Krishna et al. (2020) is a SOAP note (Podder et al., 2020) generated by their best model[8], which is based on a Pointer-Generator network (See et al., 2017); the gold reference is not provided in the paper. Although generating SOAP notes differs from our summarization task, one can see that our generated summary covers all important medical findings in the reference, with additional findings supported by the conversation (texts highlighted in yellow in Figure A.7). Our generated summary is also much more fluent than the reference paragraph in the "Miscellaneous" section of the reference. One interesting observation is the generation of *hyperlipidemia and diabetes mellitus type 2* in our summary, these findings lack direct evidence from the conversation and may be arguably hallucinations, but it is encouraging that our model successfully infers those diseases from the discussion of insulin and A1c test results in the conversation, which is a very reasonable medical connection that even human scribes are trained to do. The reference summary for the conversation from Joshi et al. (2020) (Figure A.8) is a gold reference for extractive summarization, with which our abstractive summary shows good agreement. Although some findings in our generated summary, e.g., *Her last two cycles were late by 2 weeks...*, mistakenly mixes concepts mentioned in the conversation, the summary generated by the fine-tuned model has shown promise in generalizing to a medical specialty not present in the training data (OBGYN).

5 Related work

Pretrained models Since the inception of BERT model (Devlin et al., 2018), the research community has come to the consensus that pretrained, transformer-based models can be effective zero-shot and few-shot learners and there is a constant interest to extend the generalizability and efficiency of such models. Raffel et al. (2019) studied the effectiveness of transfer learning of various transformer models and proposed a unified text-to-text framework for all text-based language tasks. Brown et al. (2020) and its earlier versions (Radford et al., 2019) showed that it is possible to elicit specific information from the model by providing an appropriate query, or "priming the model". In our work, this is effectively done in annotating each utterance with the corresponding speaker role and breaking the conversation in chunks containing information about the start of the conversation.

Long text summarization The length of input documents for summarization task is usually limited by the transformer models. One way to break this limit is to overcome the quadratic dependence on the input sequence length of attention calculation, and an abundance of novel transformer architectures with efficient attention modules have been developed in recent years, as explained comprehensively in the survey of Tay et al. (2020). Alternatively, people have been exploring hierarchical structure in the summarization models. Zhang et al. (2019a) utilized both sentence-level and document-level BERT models to hierarchically encode input documents; Grail et al. (2021) employed BERT model to encode blocks of input text followed by GRU model to integrate encodings across the blocks; Schüller et al. (2021) introduced a dynamic windowing approach for a Pointer-Generator network (See et al., 2017) to learn to shift between blocks of input as it generates summary sentences sequentially. Our multistage approach for long document summarization introduces a hierarchical structure in the training process (rather than in the model) by going from conversation snippets to a collection of incomplete (pseudo) summaries to a complete summary.

[8]The generated Assessment and Plan (A&P) section in their paper is not shown because A&P and HPI sections are largely orthogonal in content.

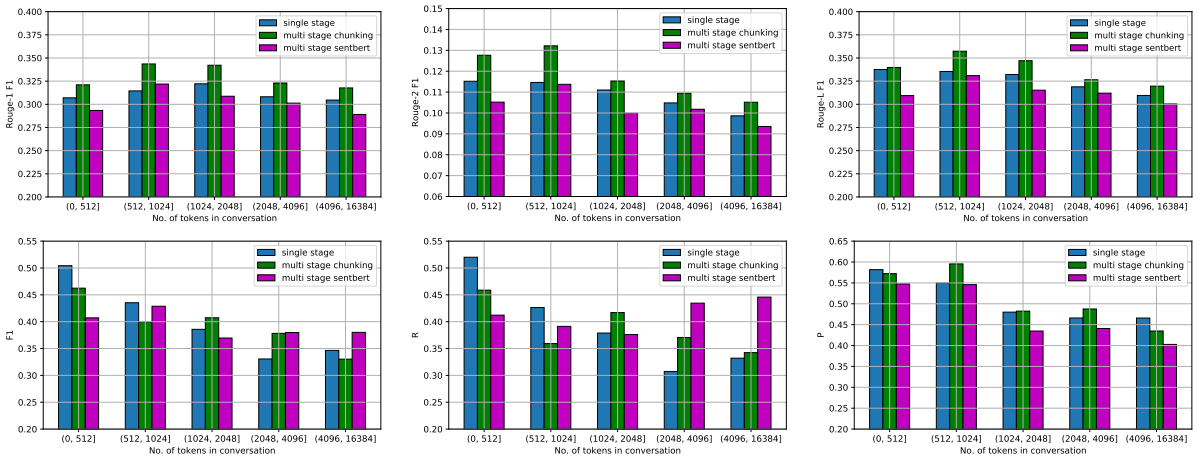

Figure 4: Performance breakdown by number of input tokens. Top row shows "mean-of-mean" ROUGE-1/2/L scores and bottom row displays concept-based F1/R/P using quickUMLS. Results for three models: single stage (blue), multistage (SentBERT) (magenta) and multistage (Chunking) (green) are shown. Vertical axis starts at nonzero for better readability.

Summarization of medical dialogue Automatic medical dialogue summarization has started to gain momentum. Krishna et al. (2020) attempted the generation of complete SOAP note from doctor-patient conversations by first extracting and clustering noteworthy utterances and then leveraging LSTM and transformer models to generate single sentence summary from each cluster. Joshi et al. (2020) showed that quality of generated summaries can be improved by encouraging copying in pointer-generator network and they also proposed alternative metrics to ROUGE for measuring the medical information coverage. There is also research to address the problems of using ROUGE for evaluating summary quality in the medical domain: Zhang et al. (2019b) explored improving factual correctness of summaries by optimizing ROUGE and concept-based metrics directly as rewards in a reinforcement learning framework of training their summarization model, although a significant difference from our work is that their task was the summarization of radiology reports instead of medical dialogues.

6 Conclusion

In this paper, we show the feasibility of summarizing doctor-patient conversation directly from transcripts without an extractive component. We fine-tune various pretrained transformer models for the task of generating the *history of present illness* (HPI) section in a typical medical report from the transcript and achieve surprisingly good performance through pretrained BART models. We propose a simple yet general two-stage fine-tuning approach for handling the input length limitation of transformer models: first, a conversation is broken into smaller portions that fit within the length budget of the model and a summarizer is trained on these portions to generate partial summaries; second, we aggregate the generated partial summaries and use them for training a second summarizer to complete the summarization. We show that this approach can help the model pick up medical findings dispersed across long conversations and reduce hallucination compared to single stage fine-tuning.

To the best of our knowledge, our work is the first to show the feasibility of generating fluent summaries directly from doctor-patient conversation transcripts. Of practical concern for medical applications, hallucination and missing information in our generated summaries can be serious problems, nevertheless, we believe our results are encouraging, especially for assisting a scribe in a human-in-the-loop system. We also plan as future work to further explore this task in the aspect of multiple reference summarization and better evaluation metrics that align with quality assessment in the medical domain.

Ethical Considerations Medical conversation summarization inevitably deals with medical data which could potentially contain sensitive information about patients and doctors alike. Careful de-identification for removing all sensitive and identifiable information in the input data is an important tool for privacy protection. We ensured that our

data went through a similar process to not reveal any sensitive information (age, name, home address, etc.) about all people involved or mentioned in the conversation. The same de-identified data is also presented to scribes during annotation to ensure no leakage of sensitive information. No information about gender, ethnicity or other discriminating factors are used as a part of our proposed method.

The intended use of our method is for designing an automatic summarization system aimed at reducing physician and scribe burnout due to the burdersome documentation process required for each medical encounter. The most natural application of this technology is not as a replacement for a human scribe, but as an assistant to one. By providing tools that aid a human scribe one can mitigate much of the risk of system failures, such as hallucination. Nonetheless, continued work is required in this area to ensure that both privacy and data accuracy are preserved.

References

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Edward Loper Bird, Steven and Ewan Klein. 2009. *Natural Language Processing with Python.* O'Reilly Media Inc.

Tom 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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Quentin Grail, Julien Perez, and Eric Gaussier. 2021. Globalizing BERT-based transformer architectures for long document summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

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

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr Summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary Lipton. 2020. Generating SOAP notes from doctor-patient conversations. *arXiv preprint arXiv:2005.01795*.

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*.

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*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2020. Soap notes. *StatPearls [Internet]*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Leon Schüller, Florian Wilhelm, Nico Kreiling, and Goran Glavaš. 2021. Windowing models for abstractive summarization of long texts. In *European Conference on Information Retrieval*, pages 384–392. Springer.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

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

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019a. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. *arXiv preprint arXiv:1905.06566*.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D Manning, and Curtis P Langlotz. 2019b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *arXiv preprint arXiv:1911.02541*.

A Appendix

A.1 Dataset Statistics

See Figure A.1 for statistics on word count and number of reference summaries in the dataset.

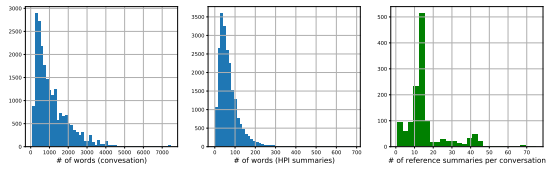

Figure A.1: Dataset statistics: word count in conversation (left); word count in HPI summaries (middle); no. of reference summaries (right).

A.2 Hyperparameters

Table A.1 lists the typical hyperparameters used in training and inference for both BART and BigBird models. BART models are trained on AWS Sagemaker instances with a single Nvidia V100 GPU (16 GB Memory); BigBird models are fine-tuned on our internal server with a single NVidia Titan X Pascal GPU (12GB memory).

A.3 Running Example

We showcase an example conversation, the corresponding reference and model generated summaries, and extracted medical findings by quickUMLS and our rule-based system in Figure A.2-A.4. These examples are referred to throughout the paper.

A.4 Inference on Out-of-dataset Examples

Figure A.7-A.8 display summaries generated on example conversations from (Krishna et al., 2020) and (Joshi et al., 2020).

A.5 Additional Evaluation Results

Hyperparameter tuning on header length. Table A.2 shows hyperparameter tuning on the percentage of header utterances retained in all conversation chunks in the multistage (Chunking) method. All percentages are measured in unit of words, i.e., for a conversation chunk of 512 words, 25% header means the header text spans 128 words, rounded up to the end of a turn in the original conversation. 128-word header is the setting used in this paper with the best ROUGE scores and least amount of inputs truncated in the second stage fine-tuning.

Development set performance. Table A.3 shows evaluation results on the development set. Most metrics are on par or slightly worse than those obtained on the test set. Although slight overfitting was observed during model fine-tuning, the comparable model performance on both development and test set indicates that reasonable performance on unseen medical conversations of similar specialty can be expected.

Gender mismatch. Roughly 30% of the model generated summaries predict the wrong patient gender. We do not penalize such a mistake in human evaluation (Section 4.3) because (a) inferring gender is not always possible solely from the conversation transcript, nor is it necessary as this information is easily attainable; (b) the model does a good job of picking up gender pronouns if they are present in the input, but this can lead to mistakes when the gender is referring to a person other than the patient; (c) correcting gender mismatch is straightforward: we experiment with adding one sentence with the correct patient gender, *The patient is a female/male*, to all model inputs and the resulting summaries predict the patient gender in 100% of the observed examples.

Parameter	BART	BigBird (RoBERTa)	BigBird (Pegasus)
Learning rate	2.5×10^{-5}	1×10^{-5}	1×10^{-4}
LR schedule	polynomial 200 steps warmup 30000 steps total	Square root decay 100 steps linear warmup 30000 steps total	Square root decay 100 steps linear warmup 30000 steps total
Batch size	1 ($\times 8$)	1	1
Optimizer	Adam	Adam	Adafactor
Dropout	0.1	0.1	0.1
Early stopping monitor	dev set NLL loss	dev set NLL loss	dev set NLL loss
Early stopping patience	3	3	3
Beam search # of hypotheses	4	5	5
Beam search maximum generation length (# of tokens)	512	256	256
Beam search length penalty	0.2	0.7	0.7

Table A.1: Hyperparameter settings. $\times 8$ in batch size setting specifies no. of updates used in gradient accumulation.

	Truncated (%)	ROUGE-1 F1	ROUGE-2 F1	ROUGE-l F1
0% header	17.5 (3.7)	0.2893 (0.4144)	0.0934 (0.1613)	0.2971 (0.3930)
25% header	**9.5 (1.7)**	**0.3227 (0.4578)**	**0.1144 (0.1991)**	**0.3302 (0.4442)**
50% header	34.4 (17.6)	0.2921 (0.4252)	0.0942 (0.1770)	0.3000 (0.4100)
75% header	45.5 (30.6)	0.2955 (0.4212)	0.0934 (0.1678)	0.3023 (0.4064)

Table A.2: Dependence of ROUGE scores on the amount of header utterances used in Multistage (Chunking) method. Second column shows the percentage of inputs > 1024 tokens (values in parentheses are for inputs > 2048 tokens) to the second stage fine-tuning. Evaluation done on dev set.

	ROUGE			quickUMLS			rule-based		
	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	F1	Precision	Recall	F1	Precision	Recall
single stage	0.3029 (0.4364)	0.1047 (0.1841)	0.3191 (0.4285)	0.3540	0.4229	0.3430	**0.4014**	0.5239	**0.5097**
multistage (Chunking)	**0.3227 (0.4578)**	**0.1144 (0.1991)**	**0.3302 (0.4442)**	**0.3922**	**0.4764**	**0.4076**	0.3829	**0.5350**	0.4877
multistage (SentBERT)	0.2997 (0.4329)	0.0997 (0.1691)	0.3098 (0.4127)	0.3665	0.4334	0.3646	0.3580	0.4731	0.4732

Table A.3: BART fine-tuning results on dev set.

Conversation Transcript

...
[DR]: Good to see you, how are you?
[PT]: Not very well.
[DR]: No?
[PT]: No, I have a bad cold.
[DR]: Oh, no.
[PT]: For the last few days I've been down.
...
[DR]: The one good thing is that the sugar remains, it's going lower.
[PT]: Good.
[DR]: It's not quite 7, it's 7.3 -
[PT]: Okay.
[DR]: From 7.4.
[PT]: Okay.
...
[PT]: I guess, like, we should take care of my eating.
[DR]: Yeah, it's your carbs. If you can really control -
[PT]: Yeah.
[DR]: Those carbs -
[PT]: Right.
...
[DR]: When did, were you, when were you sick? Was that, when did it start?
[PT]: I arrive on Friday. I have been Saturday, Sunday and, and today and yesterday.
[DR]: Yeah, the blood shows there is an infection, so I will have to give you something. You may, you may have caught it on a plane or -
[PT]: I got -
[DR]: You think so?
[PT]: I think, I, $laugh$, I got somebody who was very sick next, sitting next to me.
...
[DR]: Maybe $de-id$, you get the flu shot.
[PT]: You mean come here?
[DR]: I don't have it. So you can -
...
[PT]: Yeah. I am feeling a little bit better today but it still, I did, I feel like just lying down and sleeping.
[DR]: Oh, really? Yeah, and you're usually more energetic than that.
[PT]: Yes, I am.
[DR]: $laugh$.
...
[DR]: How is work?
[PT]: Okay, yeah, you know I enjoy that and I really -
[DR]: You're really going to keep doing that for a while?
[PT]: Until I can't, I mean, actually my contract goes for two years.
[DR]: Oh.
[PT]: Up to May 2019 and I plan to stop there. Um-hum.
[DR]: And then -
[PT]: unk -
[DR]: What do you want to -
[PT]: Do some consultancy.
[DR]: You want to $de-id$, where you're going to be?
[PT]: Either in $de-id$ here or we are exploring the possibility of doing some business in Guatemala.
...
[PT]: So, but, but basically, I don't want to go for a long time away like I do.
[DR]: Yeah.
[PT]: It's enough.
[DR]: Do you have any, do you need any refill of anything else?
[PT]: Oh, yes, uh, you know that Viagra now, is not covered by the insurance.
...

...
[PT]: What about, uh, what about, uh, Cialis?
[DR]: The same, it will be very expensive because that will be generic next year.
[PT]: So for the time being -
[DR]: Do you want to stick with Viagra or do you want me to right Cialis?
...
[PT]: Remember you gave me both?
[DR]: Yeah.
[PT]: Why don't you give me -
[DR]: I'll give you both again and -
[PT]: Both and see what happens.
...
[DR]: So, for now, I will give you the antibiotic. Do you need anything for coughing $name$, okay?
[PT]: Um, the cough, yeah, please give me something for coughing.
[DR]: I will give you something for coughing, so -
[PT]: I've been, the, the only thing I've been doing is, um, Cepacol, something like that.
[DR]: Yeah, yeah, I'll give you something stronger.
[PT]: Yeah.
...
[DR]: At least those four things, that's the Cialis, Viagra. I'll print these out so you have them.
[PT]: Okay.
[DR]: Yeah.
[PT]: Um, what I was going to ask you about, uh, several years ago, I did a colonoscopy.
[DR]: Yeah.
...
[DR]: You don't recall? Did they find anything in you?
[PT]: No.
[DR]: Nothing?
[PT]: It was okay.
...
[PT]: What, now, the other question is about, uh, the prostrate.
[DR]: Yeah.
[PT]: Is it okay?
[DR]: That one I've checked in you, it's been okay.
[PT]: Okay.
...
[DR]: Yeah, and it has the Viagra and the antibiotic and the coughs and the jarabe, so you have all of those, so -
[PT]: Oh, you gave me jarabe and the -
[DR]: Yeah, everything is there. So you just, I would take all that to Rite Aid.
[PT]: Okay, okay, very good.
...
[PT]: Yeah, I plan not to come back until May or June.
[DR]: Okay, okay, yeah. So we should -
[PT]: Over six months.

Figure A.2: Excerpt from example conversation in the dataset. Due to privacy requirement, only portions of conversation relevant to the summaries are shown. Snippets in the left panel are within the 1024 token limit and those in the right panel are beyond the limit.

Summary	
Reference (max. symp.) The patient is a male with h/o diabetes mellitus type 2 presents for follow up. He reports cold-like symptoms including cough onset 3 days ago after sitting on the plane next to someone that was sick. On review, the patient's HGB A1c was 7.3 compared to previous at 7.4. The patient states that he needs to work on eating better. The patient is requesting refills and reports that his Viagra Rx is expensive. He reports that his last colonoscopy did not show anything abnormal. The patient states that he will likely be retiring in 2 years after his contract is up and plans on doing some consulting or business in Guatemala.	**Single-stage** The patient is a female with a hx that includes diabetes mellitus type 2 who presents for f/u appointment for cold symptoms. She reports that she has been experiencing a cold that began on Friday and continues today. She notes that she feels better today, but has been lying down and sleeping. She has not received the influenza vaccine. She is interested in receiving the flu vaccine. She has a history of hyperlipidemia and hypercholesterolemia. She states that she needs a refill on her gabapentin. She denies any chest pain, shortness of breath, palpitations, dizziness, lightheadedness, trouble chewing or swallowing. She is also interested in a refill of her Viagra.
Reference (other) Male patient presents today for followup of his hyperglycemia. His hemoglobin has improved to 7.3 to 7.4. He reports being sick after he recently traveled via plane. He notes that he has been sick for the past 3 days. He is feeling a bit better today, but still feels tired and sleepy. He says that he has been taking Cepacol for the symptoms.	**Multistage (Chunking)** The patient presents to the clinic today for a follow-up visit. He has a history of diabetes mellitus. His A1c today is 7.3. He reports that he has been experiencing a cold for the past few days. He states that he is still feeling fatigued. He is not eating as well as he would like. He would like a refill of his Cialis and Viagra. He is currently working for a consultancy and has a contract that lasts until May 2019. He will be out of the country for 6 months. He does not want to go for a long time away.
Reference (other) The patient is a ** y/o male presenting for a follow up. He reports having a "bad cold" for the past 2-3 days. He reports being fatigue and cough. He has not received his influenza vaccine. He has asking about Viagra and Cialis prescriptions.	**Multistage (Sentbert)** The patient presents to the clinic today for a follow up visit concerning a cold and cough. The patient has been complaining of a cold for the past few days. She reports that she is feeling a little better today, but she still feels like lying down and sleeping. The patient reports that her blood sugar has been down for the last few days, and she has been trying to control her diet. She has not been taking her medications as prescribed. She is requesting a refill on her Viagra, Cialis, and Cepacol for her cough. She also requests a refill of her antibiotic, cough syrup, and cough drops. She states that she has not received the flu vaccine yet. She has had a colonoscopy in the past and it was normal. She would like to get one again. She does not want to go for a long period of time away from her current medication. She will continue to work until May 2019 and then she will stop. Her husband was expecting her to do more work, so she is going to keep doing it for a while.

Figure A.3: BART generated summaries and references. Text in green highlights medical findings present in at least one reference summary; text with yellow highlighting shows findings not in reference but are supported by the conversation.

Summary	quickUMLS	In-house NLP
Reference (max. symp.): The patient is a male with h/o diabetes mellitus type 2 presents for follow up. He reports cold-like symptoms including cough onset 3 days ago after sitting on the plane next to someone that was sick. On review, the patient's HGB A1c was 7.3 compared to previous at 7.4. The patient states that he needs to work on eating better. The patient is requesting refills and reports that his Viagra Rx is expensive. He reports that his last colonoscopy did not show anything abnormal. The patient states that he will likely be retiring in 2 years after his contract is up and plans on doing some consulting or business in Guatemala.	diabetes mellitus type 2, diabetes mellitus type 1a, diabetes mellitus type 1b, patient state, colonoscopy, consulting, symptoms, abnormal, retiring, business, contract, present, report, sitting, <2 years, review, eating, Viagra, like, cough, cold, hgb	finding reported by subject or history provider cough does sit sitting position retired (life event)
Single-stage: The patient is a female with a hx that includes diabetes mellitus type 2 who presents for f/u appointment for cold symptoms. She reports that she has been experiencing a cold that began on Friday and continues today. She notes that she feels better today, but has been lying down and sleeping. She has not received the influenza vaccine. She is interested in receiving the flu vaccine. She has a history of hyperlipidemia and hypercholesterolemia. She states that she needs a refill on her gabapentin. She denies any chest pain, shortness of breath, palpitations, dizziness, lightheadedness, trouble chewing or swallowing. She is also interested in a refill of her Viagra.	diabetes mellitus type 2, diabetes mellitus type 1a, diabetes mellitus type 1b, hypercholesterolemia, shortness of breath, flu vaccine, dizziness, lightheadedness, hyperlipidemia, cold symptoms, poor swallowing, palpitations, interested, gabapentin, chest pain, sleeping, present, report, history, chewing, viagra, notes, feels, cold	finding reported by subject or history provider patient feels better does lie down interested pain chest pain shortness of breath palpitations dizziness lightheadedness does chew does swallow interested
Multistage (Chunking): The patient presents to the clinic today for a follow-up visit. He has a history of diabetes mellitus. His A1c today is 7.3. He reports that he has been experiencing a cold for the past few days. He states that he is still feeling fatigued. He is not eating as well as he would like. He would like a refill of his Cialis and Viagra. He is currently working for a consultancy and has a contract that lasts until May 2019. He will be out of the country for 6 months. He does not want to go for a long time away.	history of diabetes mellitus, follow-up visit, ill feeling, fatigued, contract, present, report, country, eating, cialis, viagra, cold, like, may	fatigue
Multistage (Sentbert): The patient presents to the clinic today for a follow up visit concerning a cold and cough. The patient has been complaining of a cold for the past few days. She reports that she is feeling a little better today, but she still feels like lying down and sleeping. The patient reports that her blood sugar has been down for the last few days, and she has been trying to control her diet. She has not been taking her medications as prescribed. She is requesting a refill on her Viagra, Cialis, and Cepacol for her cough. She also requests a refill of her antibiotic, cough syrup, and cough drops. She states that she has not received the flu vaccine yet. She has had a colonoscopy in the past and it was normal. She would like to get one again. She does not want to go for a long period of time away from her current medication. She will continue to work until May 2019 and then she will stop. Her husband was expecting her to do more work, so she is going to keep doing it for a while.	taking medication, current medication, follow up visit, patient reports, a little better, cough syrup, cough drops, flu vaccine, colonoscopy, prescribed, antibiotic, sleeping, present, request, report, feels, cepacol, controll, viagra, cialis, normal, period, blood, sugar, cough, cold, like, diet,	cough does lie down

Figure A.4: Medical concepts extracted from summaries by quickUMLS and our rule-based system. UMLS findings (second column) are separated by commas, and rule-based findings (third column) are shown on separate lines. In order to control the generation of false positive concepts, we choose to consider for evaluation only clinical findings (symptoms) extracted by the In-house NLP system; disorders (e.g. diabetes mellitus), medications and clinical procedures (e.g., colonoscopy) are ignored, which are concepts of lower priority in the HPI section of an EHR report.

Conversation Chunks

2nd Chunk:

[PT]: Hi. [DR]: Hey. [PT]: How are you? [DR]: Good to see you, how are you? [PT]: Not very well. [DR]: No? [PT]: No, I have a bad cold. [DR]: Oh, no. [PT]: For the last few days I've been down. [DR]: Oh, no. [PT]: unk down. [DR]: We should try to take care of that problem then. [PT]: Please do. [DR]: The one good thing is that the sugar remains, it's going lower. [PT]: Good. [DR]: It's not quite 7, it's 7.3 - [PT]: Okay. [DR]: From 7.4. [PT]: Okay. [DR]: So, at least you're going in the right - [PT]: Direction. [DR]: Direction, so - [PT]: Uh - [DR]: We just continue - [PT]: I guess, like, we should take care of my eating. [DR]: Yeah, it's your carbs. If you can really control - [PT]: Yeah. [DR]: Those carbs - [PT]: Right. [DR]: That will help a lot. So - [PT]: Yeah. **...** [PT]: I, $laugh$, yeah. My poor husband is very, and he was, uh, he was expecting for me to come here and do things and I didn't unk - [DR]: Oh, no. [PT]: Doing nothing. [DR]: How is work? [PT]: Okay, yeah, you know I enjoy that and I really - [DR]: You're really going to keep doing that for a while? [PT]: Until I can't, I mean, actually my contract goes for two years. [DR]: Oh. [PT]: Up to May 2019 and I plan to stop there. Um-hum. [DR]: And then - [PT]: unk - [DR]: What do you want to - [PT]: Do some consultancy. [DR]: You want to $de-id$, where you're going to be? [PT]: Either in $de-id$ here or we are exploring the possibility of doing some business in Guatemala. [DR]: Oh. [PT]: Together with $name$. [DR]: Yeah, yeah. [PT]: So, but, but basically, I don't want to go for a long time away like I do. [DR]: Yeah. [PT]: It's enough. [DR]: Do you have any, do you need any refill of anything else? [PT]: Oh, yes, uh, you know that Viagra now, is not covered by the insurance. [DR]: Oh. [PT]: $90 per pill. [DR]: Because at the end of the year it will become generic. [PT]: That's what I heard. [DR]: But right now, it's extremely expensive. [PT]: Right. [DR]: One pill, I have somebody $de-id$, bought it for $70. [PT]: I, I mean - [DR]: Um - [PT]: CVS told me $90. [DR]: Oh, my gosh, you have to shop around, um - [PT]: What about, uh, what about, uh, Cialis? [DR]: The same, it will be very expensive because that will be generic next year. [PT]: So for the time being - [DR]: Do you want to stick with Viagra or do you want me to right Cialis? [PT]: I, I don't know - [DR]: They're both - [PT]: The price, I mean - [DR]: They' both will be very expensive. [PT]: Remember you gave me both? [DR]: Yeah. [PT]: Why don't you give me - [DR]: I'll give you both again and - [PT]: Both and see what happens. [DR]: And see what, whether are the cheaper one is bad. They're both going to be very expensive because, $laugh$. [PT]: Wow. [DR]: When things go generic, the drug companies try to really, uh, get your money out of you before. [PT]: Yeah, yeah, the drug companies are such thief. [DR]: So, for now, I will give you the antibiotic. Do you need anything for coughing $name$, okay? **...**

Last chunk:

[PT]: Hi. [DR]: Hey. [PT]: How are you? [DR]: Good to see you, how are you? [PT]: Not very well. [DR]: No? [PT]: No, I have a bad cold. [DR]: Oh, no. [PT]: For the last few days I've been down. [DR]: Oh, no. [PT]: unk down. [DR]: We should try to take care of that problem then. [PT]: Please do. [DR]: The one good thing is that the sugar remains, it's going lower. [PT]: Good. [DR]: It's not quite 7, it's 7.3 - [PT]: Okay. [DR]: From 7.4. [PT]: Okay. [DR]: So, at least you're going in the right - [PT]: Direction. [DR]: Direction, so - [PT]: Uh - [DR]: We just continue - [PT]: I guess, like, we should take care of my eating. [DR]: Yeah, it's your carbs. If you can really control - [PT]: Yeah. [DR]: Those carbs - [PT]: Right. [DR]: That will help a lot. So - [PT]: Yeah. **...** [DR]: But I didn't see any of that. It was just really interesting - [PT]: A nice city - [DR]: $de-id$, I like the city - [PT]: Oh, yeah. [DR]: It's nice, it's nice city. [PT]: Yeah. [DR]: But Catalani just don't understand, right? The language is very different. [PT]: Yeah, but everyone, uh, speak Spanish, yeah. [DR]: Yeah. [PT]: Yeah, Catalani is different. It's a unk. [DR]: That's different. [PT]: Yeah. [DR]: And then we want to $de-id$, just like further up a little bit. That's very Catalonian, they know - [PT]: Oh, right. [DR]: Yeah. [PT]: Maybe, yeah. [DR]: I went to a museum there and they did not really - [PT]: Yeah, yeah, I have a good friend - [DR]: They speak Spanish - [PT]: And I went with $name$ to, to her wedding there. She's a Catalani. [DR]: Where? In - [PT]: In $de-id$. [DR]: Yeah? [PT]: Yeah. [DR]: Nice city, right? [PT]: Yeah, very beautiful city. [DR]: Nice city. Uh, so - [PT]: And very good food. [DR]: Yeah, yeah, they - [PT]: I remember - [DR]: They eat very, you know, at the very last meal, we had dinner at 11:00. That's crazy. Those people eat, $laugh$. [PT]: That's my culture. [DR]: They eat that late? [PT]: Yeah, in Chile, we, the regular time for eating in Chile is 9:00 PM, never before 9:00. [DR]: Really? [PT]: But we can keep going later. [DR]: What time do you sleep? [PT]: Uh, we don't get up early. [DR]: $laugh$. [PT]: We don't like getting up early, $laugh$. [DR]: I found out though - [PT]: Yeah. [DR]: Here they have, uh, the siesta, where they close down the shops too and they - [PT]: Right, they do have - [DR]: Yeah. [PT]: They do that. [DR]: Do the doctors do that too? They take a break? [PT]: I don't know. [DR]: $laugh$. [PT]: I don't know. Uh - [DR]: I was like, do doctors - [PT]: Maybe you are considering going there to - [DR]: Oh, I don't know, that sounds - [PT]: To practice unk, $laugh$. [DR]: Like a good life, $laugh$. [PT]: Yeah. [DR]: Good to see you, feel better, okay? [PT]: Okay, thank you. [DR]: Drink lots of water. [PT]: Thank you. [DR]: Good to see you. [PT]: Nice to see you. [DR]: Bye bye. Great to see you. [PT]: Bye bye now. [DR]: Bye bye.

Summary

The patient is a ** y/o male, presenting for a follow up. He has a cough and cold. He reports that he has not been feeling well for the past few days. He has been coughing up phlegm. His blood sugar has been lower. He is compliant with his medications. He would like a refill of his Viagra and Cialis. He does not want to go for a long time away.

The patient is a ** y/o male presenting for a follow up. He has a history of diabetes mellitus. His A1c is 7.3. He is compliant with his medications. He reports that he has been experiencing a cold for the past few days. He has not been eating as well as he would like. He is considering going to Chile.

Figure A.5: Example conversation chunks and generated summary in the first stage of multistage fine-tuning with Chunking method. Text highlighted in green are sentences common to both summaries. Text in yellow marks sentences that are supported by the "body" part of each chunk. Note that ... is used in each chunk to mask the rest of the conversation.

Conversation Snippets	Single Sentence Summary
[PT]: Hi. [DR]: Hey. [PT]: How are you? [DR]: Good to see you, how are you? [PT]: Not very well. [DR]: No? [PT]: No, I have a bad cold. [DR]: Oh, no.	Patient states that she has a bad cold.
[PT]: Right. [DR]: One pill, I have somebody $de-id$, bought it for $70. [PT]: I, I mean – [DR]: Um – [PT]: CVS told me $90. [DR]: Oh, my gosh, you have to shop around, um – [PT]: What about, uh, what about, uh, Cialis? [DR]: The same, it will be very expensive because that will be generic next year.	Patient reports that he has been taking Cialis for erectile dysfunction and states that it is expensive.
[PT]: Um, what I was going to ask you about, uh, several years ago, I did a colonoscopy. [DR]: Yeah. [PT]: Uh, it's not that I want to do it again, but do you think I should do it again? [DR]: They will give an indication of how often they want to see you after that initial one. Do you remember if they send the records here or was it somewhere else? Do you know if they sent your records to me? I'm not seeing it in my records. [PT]: Oh, it wasn't to you. [DR]: It wasn't to me? [PT]: No, it was my previous doctor. [DR]: Remember what, um, because the letter will say, we want to see $name$ back in 3 or 5 or 10 years. There's a number.	Patient states that he had a colonoscopy several years ago and would like to know if he should do it again.
[PT]: Yeah, I'll be here almost entire month of, uh, December. [DR]: I'm here in December. [PT]: Okay, good. [DR]: Why don't you come back then and then, after, you'll be away again after that, right? [PT]: Yeah. [DR]: Yeah, yeah. [PT]: Yeah, I, I plan to see you, because then I will be out for a while. [DR]: Oh, you are? Okay.	Patient states that he will be out of the country for a month in December.
[PT]: Yeah. [DR]: Good to see you, feel better, okay? [PT]: Okay, thank you. [DR]: Drink lots of water. [PT]: Thank you. [DR]: Good to see you. [PT]: Nice to see you. [DR]: Bye bye. Great to see you.	Patient is doing well overall and states that he is drinking lots of water.

Figure A.6: Example conversation snippets and single sentence summary used in the first stage of multistage fine-tuning with the SentBERT method. Snippets chosen approximately equi-distance from each other from the beginning to the end of the conversation. Text highlighted in green show generated contents and the supported text in the corresponding snippets.

Conversation Transcript

[DR]: Okay, so, um, we are going to talk a little bit about being a basal insulin candidate .
[DR]: Um, we have talked about your A1c and the things, what are, so what are the things that , that keep you from, um , from being the best possible diabetic that you possibly can ?
[DR]:, I know there's a lot of stuff that troubles you.
[PT]: Snacking and stress eating.
[PT]: Eating late in the evenings instead of, um, at a reasonable time -
[DR]: Right.
[PT]: At night, late.
[PT]: Poor meal planning.
[DR]: Right, and I think that's in the, we can all take a little note for but one of things that really got me worried because your last A1c was really high -
[PT]: Uh-huh.
[DR]: It was above , it was above 10 , and we 've had this consistent pattern and you 've really , I mean , you really have given it an effort and I have to give it up to you that you 've been trying and , um , so we 're down to like just a couple of options and so I want to just kind of put them before you .
[DR]: I 've got, I 'm, I 'm considering once a day insulin with you at some point .
[DR]: Um, I do n't want to use that as a threat.
[DR]: I do n't want to use it as like a, oh , you 've been a bad patient you deserve to be on basal insulin .
[DR]: Um , I do have one other option , um , but I want to counsel you that , that basal insulin , even if , if we did , we do go to it , it is not a punishment .
[DR]: It is something to kind of get your baseline down to a regular, regular situation and you only have to do it once a day.
[DR]: Um, and I know that one of the things that we have for diabetics is their eating habits .
[DR]: And, so , I am proposing as instead of using insulin this time , um , that we use something called Vyvanse for the , for the eating at nighttime .
[DR]: Um, it's supposed to reduce the incidence of having those nighttime cravings so that you can work , you can do your things , you can plan a little bit better .
[DR]: It 's , it's originally for ADHD so some people actually feel a little bit more focused , um , and controlled but it also affects appetite centers and so it's supposed to do it for the longer term as opposed to using like a fen phen , um , so , which is short term .
[DR]: So, um , I 'm really hoping with your interest in it and with the coverage hopefully , I know , with your particular plan it should be covered and we can get a discount .
[DR]: Um, we do it once a day with your other medications , which are actually pretty minor .
[DR]: Um, and then we check you again in eight weeks .
[PT]: Okay.
[DR]: All right?
[DR]: And, so what we do is we say , you know , it should be , we usually will do three months but then eight weeks we should see some difference from today .
[DR]: We should see some kind of improvement and then we can sort of celebrate that in and of itself, if that's okay with you.
[PT]: That sounds great.
[DR]: Cool, all right well we will plan to meet again in eight weeks .
[PT]: Okay.
[DR]: And, uh , and we 'll go from there .
[PT]: Okay.
[DR]: Cool, all right , cool .
[PT]: Thank you.

Summaries

Krishna et al (2020):

Chief Complaint: basal insulin .
Review of Systems: snacking and stress eating .
Past Medical History: high a1c .
Medications: insulin .
Miscellaneous: patient eating late in the evenings instead of a reasonable time at night . last a1c was really high above , it was above , it was above , it was above , it was above , it was above , it was above , it was above , it was above 10 , and had this consistent pattern and the consistent pattern , you really have given it an effort and i have to give it up to you that you 've been trying , so we 're down to just a couple of options and so i want to just kind of put them before .

Our single-stage model:

The patient is a ** y/o male presenting for follow up for diabetes.
He has a history of diabetes mellitus type 2 and hyperlipidemia.
He reports that he has been trying to watch his diet.
He admits to eating late at night and having poor meal planning.
He states that his last A1c was above 10.
He would like to start on basal insulin.
He is interested in using Vyvanse to reduce his nighttime cravings.
He has no other concerns at this time.

Figure A.7: Inference on example conversation in (Krishna et al., 2020). Text with green highlighting shows medical findings common in both summaries; text with yellow highlighting shows findings unique to our single-stage model generation that are supported by the conversation. We choose to omit the Assessment&Plan section from the original paper as HPI and Assessment&Plan have little overlap in contents in a medical report.

Conversation Transcript

[DR]: what was your last period date ?
[PT]: feb 9th to 13th I have an average cycle of about 32-33 days.
[DR]: are/were you on any hormonal form of birth control apart from plan b?
[PT]: no
[DR]: Thanks for letting me know.
[DR]: how regular are your cycles usually?
[PT]: they can sometimes be off by a couple days, give or take because i have hypothyroidism and am taking synthroid. but as of lately with my last two cycles, they had predicted to the day or a day late.
[DR]: okay . Is this the first time you are missing period this late?
[PT]: no . Ive had it be late by two weeks and even have missed it twice.
[DR]: okay. Have you been trying to lose weight?
[PT]: Ive been watching what Ive been eating, so yes .
[DR]: any recent change in your physical activity?
[PT]: no
[DR]: when was the last time you had your thyroid panel checked/tested ?
[PT]: just last week. everything is as normal as can be.
[DR]: that's great to know.

Summaries

Reference:

last menstrual period is february 9th-13th. has average cycles of 32-33 days.
not on hormonal form of birth control apart from plan b.
cycles may be off by a couple days because of hypothyroidism and is taking synthyroid. since last 2 cycles they are predicted to the day or a day late.
not the first time missing period. has it late by 2 weeks and even missed it twice.
is watching what he or she eats to loose weight.
no recent changes in physical activity.
checked thyroid panel last week and everything is normal.

Joshi et al., 2020:

period date feb 9th to 13th. average cycle of about 32-33 days
no hormonal form of birth control apart from plan b.
they can sometimes be off by a couple day. has hypothyroidism and am taking synthroid. has predicted to the day or a day late.
had it be late by two weeks and even have missed twice.
trying to lose weight. been watching what is eating, so yes.
no recent change in physical activity
had thyroid panel checked/tested just last week. everything is as normal

Our single-stage model:

The patient is a ** y/o female presenting for f/u for hypothyroidism.
She reports that her cycles can sometimes be off by a couple days because she has hypothy thyroidism and is taking synthroid.
She has an average cycle of 32-33 days.
Her last two cycles were late by 2 weeks and she has missed her period twice.
She denies any recent change in her physical activity.
She is not on any hormonal form of birth control.
She had her thyroid panel checked last week.

Figure A.8: Inference on example conversation from (Joshi et al., 2020). Text with green highlighting shows medical findings common in both summaries.