An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters

Asma Ben Abacha
Microsoft Health AI
abenabacha@microsoft.com

Wen-wai Yim
Microsoft Health AI
yimwenwai@microsoft.com

Yadan Fan
Nuance Communications
yadan.fan@nuance.com

Thomas Lin
Microsoft Health AI
tlin@microsoft.com

Abstract

Medical doctors spend on average 52 to 102 minutes per day writing clinical notes from their patient encounters (Hripcsak et al., 2011). Reducing this workload calls for relevant and efficient summarization methods. In this paper, we introduce new resources and empirical investigations for the automatic summarization of doctor-patient conversations in a clinical setting. In particular, we introduce the MTS-DIALOG dataset; a new collection of 1,700 doctor-patient dialogues and corresponding clinical notes. We use this new dataset to investigate the feasibility of this task and the relevance of existing language models, data augmentation, and guided summarization techniques. We compare standard evaluation metrics based on n-gram matching, contextual embeddings, and Fact Extraction to assess the accuracy and the factual consistency of the generated summaries. To ground these results, we perform an expert-based evaluation using relevant natural language generation criteria and task-specific criteria such as critical omissions, and study the correlation between the automatic metrics and expert judgments. To the best of our knowledge, this study is the first attempt to introduce an open dataset of doctor-patient conversations and clinical notes, with detailed automated and manual evaluations of clinical note generation.

1 Introduction

The recent progress in automatic summarization has been highly influenced by large transformer-based language models and the availability of large-scale datasets. The summarization of medical conversations is well positioned to benefit from similar approaches, but is facing domain- and task-specific obstacles such as the lack of data and relevant evaluation protocols.

Medical doctors spend on average 52 to 102 minutes per day writing clinical notes from their conversations with the patients (Hripcsak et al., 2011). The time spent with Electronic Health Record systems contributes to work-life imbalance, dissatisfaction, high rates of attrition, and a burnout rate exceeding 50% (Arndt et al., 2017). Summarization models could play a key role in reducing that workload by generating clinical notes from the doctor-patient encounters (Knoll et al., 2022).

Summarizing doctor-patient conversations for a clinical setting brings its own set of challenges and nuances in addition to the typical natural language understanding and generation components. For instance, omission of critical medical facts is likely to alter patient outcomes and should be one of the critical/deciding factors in adopting one summarization system over another. Hallucinations are also likely to impact the clinical outcome if they are not avoided (or detected with a high accuracy).

Designing and improving summarization models that address these challenges can benefit from wider research efforts on the task. However, the lack of publicly available doctor-patient dialogue datasets limits wider research efforts from the NLP community in this summarization task. In this paper, we tackle the lack of data for the task by building a new dataset of doctor-patient conversations and associated clinical notes. We avoid privacy infringement risks, by creating simulated conversations from publicly available clinical notes.

Our main contributions are: **(i)** a new dataset of 1,700 doctor-patient conversations (16k turns and 18k sentences) and their summarized clinical notes (6k sentences). To the best of our knowledge, this is the first publicly-available dataset of medical conversations and associated notes at this scale, **(ii)** an evaluation of several SOTA summarization models, including variants that use existing relevant datasets, augmented training data, and guided summarization for medical conversation summarization, and **(iii)** a study of standard evaluation metrics, domain-specific metrics, and expert judgments for the task, including computing the corre-

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2291–2302

lation between the automatic and manual scores for the evaluation of the generated clinical notes[1].

2 Related Work

Summarization datasets and evaluation methods are often centered around large news articles such as the CNN/DailyMail dataset of 313k newspaper articles (Hermann et al., 2015) and XSum (Narayan et al., 2018) with 227k BBC articles with single-sentence summaries.

Dialogue summarization is relatively less studied in open domain, with a few efforts tackling the summarization of conversations and meetings (Goo and Chen, 2018; Li et al., 2019; Shin et al., 2022) and the creation of meeting summarization datasets (Janin et al., 2003; Carletta, 2007).

Medical conversation summarization is also under-studied, except for a few recent efforts (Liu et al., 2019; Kazi and Kahanda, 2019; Yim and Yetisgen, 2021; Michalopoulos et al., 2022). For instance, (Joshi et al., 2020) applied a pointer generator model to generate summaries from doctor-patient dialogues in telemedicine. Instead of modeling the whole dialogue, they trained the model on snippets taken from the dialogue turns. (Zhang et al., 2021) fined-tuned a pre-trained BART model to automatically generate summaries from doctor-patient conversations. However, they only focused on two specialties (internal medicine and primary care) and the training data only consist of HPI (History of Present Illness) section. (Krishna et al., 2021) proposed an algorithm called Cluster2Sent to generate SOAP notes from doctor-patient conversations. Their summarization model involves both abstractive and extractive methods. (Enarvi et al., 2020) studied both RNN and transformer-based sequence-to-sequence models for generating medical reports. They experimented on Orthopedics data and found that sequence-to-sequence modeling is more promising. The datasets of the studies mentioned above are not made public.

(Song et al., 2020) investigated medical conversation summarization from a Chinese conversational corpus. They applied a hierarchical encoder-tagger model for extractive summarization to generate two types of summaries; one for problem statement and one for the treatment recommendations. Several other medical conversation datasets have been recently created from Chinese online health platform conversation (Liu et al., 2020; Zhang et al., 2020; Lin et al., 2019) but they do not include conversation summaries.

(Moramarco et al., 2022b) studied the task of consultation note generation on a small set of 57 transcript-note pairs (Papadopoulos Korfiatis et al., 2022) and performed a correlation study with several automatic metrics. As their focus was to compare automated metrics with human judgements, they chose models that would "produce different outputs to cover a wider range of errors" instead of attempting to benchmark SOTA summarization models. From the set of metrics they tested, they noted that character-based Levenshtein distance, BERTScore, and METEOR performed best for evaluating the note generation task.

3 MTS-Dialog

3.1 Data Creation

Our data creation approach consists in generating simulated doctor-patient conversations from publicly available clinical notes. To gather these clinical notes/summaries, we collected notes from the public Mtsamples collection, which provides de-identified clinical notes[2] (South et al., 2014; Moramarco et al., 2022a). The selected clinical notes cover the six most frequent note types and specialties in the collection, including: General Medicine, SOAP (Subjective, Objective, Assessment, Plan), Neurology, Orthopedic, Dermatology, and Allergy/Immunology.

Eight trained annotators, with medical backgrounds, were given all the sections from the clinical notes and asked to create clinical conversations from one section at a time according to detailed guidelines. These guidelines were developed based on an analysis of a large private collection of hundreds of real doctor-patient conversations and associated notes. The annotation guidelines included:

- Conversation creation rules: conversations should be (i) written as it pertains to the day of the visit with the patient and (ii) framed in the context of either an outpatient visit or emergency room visit.
- Medical terms rules: The clinical note may have a more detailed referring expression to the same problem, treatment, or test, than

[1]The dataset, source code, and annotations are available at https://github.com/abachaa/MTS-Dialog

[2]www.mtsamples.com. Mtsamples categorizes notes to at least one of 40 specialty/note-type labels. Each note explicitly marks section headers as bolded HTML tags.

Dialogue:
Doctor: My chart here says that you're eighty three years old, is that correct, ma'am?
Patient: Yes doctor, that's correct, I just had my birthday.
Doctor: Happy belated birthday! How have you been doing since your last visit?
Patient: Well, my cancer hasn't needed phlebotomies for several months now, which is good.
Doctor: That's great, you have been treated for polycythemia vera, correct?
Patient: Yes, that's the one.
Doctor: I also see you're unassisted today, which is also great.
Patient: Yeah, having some independence is nice.
Section header: History of Present Illness
Section text: The patient is an 83-year-old female with a history of polycythemia vera. She comes in to clinic today for followup. She has not required phlebotomies for several months. The patient comes to clinic unaccompanied.

Table 1: Example of a doctor-patient conversation and associated note/summary from the MTS-DIALOG dataset.

what would be represented in the conversation; for example "Open reduction internal fixation (ORIF)" from the clinical note could be translated to "We will have to do surgery on it" in the conversation.

- Imaginary but plausible rule: if the clinical note is underspecified, it is possible to create a conversation as long as it is plausible. However, the dialogues should be created so that the transcripts are more detailed than the associated clinical notes (except for problems, treatments and tests as mentioned above).
- Formatting rules: annotators should follow standard transcript guidelines such as writing words as they would be pronounced, and capitalization and punctuation rules.
- Conversation characteristics: the goal is to create a dataset with as much variation as possible to mimic real doctor-patient medical visits, including the use of speech disfluencies such as false starts, filler words, interjections, interrupting speech, corrections by the speaker to previous information, using slang and vernacular, and colloquial terms.

We also normalized the 279 original section headers from the notes into 20 types of first-level headers (e.g. *assessment, allergy, diagnosis, exam, medications, past medical history, past surgical*).

The final MTS-DIALOG dataset includes 1,701 pairs of dialogues and associated sections from the clinical notes. Table 1 presents an example from the dataset. The number of pairs from the respective specialties and note types were General Medicine:1,035, SOAP:79, Neurology:296, Orthopedic:208, Dermatology:56, and Allergy/Immunology:27. The dataset was created over a cumulative total of approximately 1,800 hours. Additional statistics are shown in Table 2.

	Dialogue			**Summary**	
	Turns	Sentences	Words	Sentences	Words
count	15,969	18,406	241,685	5,870	81,299
mean	9	11	142	3	48
max	103	136	1,951	57	1,182
25-perc	4	4	48	1	6
50-perc	6	7	88	2	18
75-perc	12	14	176	4	55

Table 2: Statistics of the MTS-DIALOG Dataset.

3.2 Data Quality

The quality of the MTS-DIALOG dataset is ensured by three stages: (1) only candidates with medical training were hired as annotators (e.g. former medical scribes), and (2) the training for this task involved one-on-one periodic feedback during initial stages by an experienced trainer; (3) at the completion of the entire dataset, a separate and independent validation step was conducted to formally evaluate the corpus against a rubric grading system. This independent evaluation graded the conversations according to their adherence with the annotation guidelines and content relevance and coverage w.r.t. the initial clinical note. Table 3 presents the results of this manual validation. The validator was additionally tasked to perform minor corrections (e.g. misspellings or adding back missed information), ensuring that the final data quality would be higher than those reported.

3.3 Comparison with Real Data

The MTS-DIALOG dataset consists of real notes and synthetic conversations that simulate doctor-patient encounters to avoid the public release of private doctor-patient conversations. To study the impact of relying on synthetic data, we investigated the resemblance of the MTS-DIALOG data with real conversations through a blind evaluation of two equal subsets of 52 conversations randomly extracted from (i) the MTS-DIALOG dataset and (ii) a private collection of recorded and tran-

Score	Description	Freq (%)
0.1	Unviable, content is not covered.	25 (1%)
0.3	Content is covered but with logical and medical errors OR Major content is not covered OR At least one major socio-cultural dialogue discrepancy anomaly.	189 (11%)
0.5	Acceptable but some minor content or sociocultural discrepancy issues.	70 (4%)
0.7	Acceptable but with misspellings or transcription rules errors only.	551 (32%)
1.0	Follows guidelines completely, logically and medically sound, no content errors or other issues.	866 (51%)

Table 3: Data validation scoring rubric and frequencies based on manual validation, prior to correction.

scribed doctor-patient conversations. Turns from the real/private subset were selected to match the length of the MTS-DIALOG turns. A medical expert with experience working as a medical scribe in hospitals independently performed this blind annotation by labeling each conversation as real, synthetic, or unknown, together with a written explanation behind each annotation. The annotation criteria included annotating the characteristics of the conversations in terms of disfluency and interruptions.

Label	#Ref	#LabeledAs	TP	FP	FN	P	R	F1
Real	52	34	29	5	23	0.85	0.56	0.67
Synthetic	52	69	47	22	5	0.68	0.90	0.78
Unknown	0	1	0	1	0	0	1	0
Total	104	104	76	28	28	0.73	0.73	0.73

Table 4: Blind Evaluation: #Ref (number of reference samples), #LabeledAs (number of assigned labels), TP (True Positive), FP (False Positive), FN (False Negative), P (Precision), R (Recall), and F1 Score.

The results of this annotation (cf. Table 4) show that 55.77% of real conversations (29/52) were annotated as real, 42.31% of real conversations (22/52) were labeled as synthetic, and 9.61% of synthetic data (5/52) labeled as real. The medical expert's blind labeling was incorrect 26.92% of the time (28/104). If synthetic and real were indistinguishable, the incorrect rate would be 50%.

A common explanation given by the medical expert for cases where synthetic data was labeled as real is that the content seemed realistic, despite the statistical comparison in Table 5 which shows that the MTS-DIALOG conversations had less disfluencies and interruptions on average. Common explanations for labeling real data as synthetic included the conversation being "to-the-point", clear with low disfluencies, easy to follow, containing abrupt subject changes and containing colloquial speech.

This difficulty in distinguishing synthetic from real indicates that the MTS-DIALOG dataset is a valuable initial dataset, for use in training and benchmarking models for real-world applications, including using the MTS-DIALOG data for data augmentation or as pre-training data for later fine-tuning on (private) real conversations data.

Dataset	#Turns	MaxTurnLen	#Disfluency	#Interruptions
MTS	7.10 (369)	18.54 (n/a)	0.63 (33)	1.06 (55)
Real	7.10 (369)	34.00 (n/a)	1.77 (92)	1.98 (103)

(a) Statistics on the full annotated datasets

Dataset [#Conversations]	#Turns	MaxTurnLen	#Disfluency	#Interruptions
MTS-mistaken-as-Real [5]	5.0	11.2	0.4	1.0
Real-mistaken-as-MTS [22]	5.9	27.2	1.4	1.5

(b) Statistics on the incorrectly labeled subsets

Table 5: Blind Evaluation: Statistics (Mean (Sum))

4 Methods

4.1 Summarization Models

To study the specificity of doctor-patient conversation summarization and the relevance of evaluation methods, we generated summaries using several SOTA transformer-based models (e.g. BART (Lewis et al., 2020) and Pegasus (Zhang et al., 2020)), including variants that are pre-finetuned using relevant datasets (e.g. XSum (Narayan et al., 2018) and Samsum (Gliwa et al., 2019)), as well as augmented training data and guided summarization as described in the following sections.

4.2 Data Augmentation via Back-translation

Relevant data augmentation is an effective technique to avoid over-fitting and increase the performance of neural methods. In particular, back-translation augmentation consists in translating the text to another language and then translating it back to the original language. We used two different auxiliary languages to add more variety in the back-translated sentences[3]. To reduce translation errors, we selected French and Spanish for their lexical proximity with English, and high-performing translation models (Tiedemann and Thottingal, 2020).

4.3 Guided Summarization

Several guidance signals can be used to control the output of summarization models. A clinical note

[3]We also release the augmented dataset of 3,603 pairs of medical conversations and associated notes.

consists of several sections such as Family History and Assessment, and the simulated conversations from the MTS-DIALOG dataset were generated independently for each section. In the guided summarization trial, we used the section header as a prefix in the training data to guide the summarization of the doctor-patient conversation. This allows the model to learn to generate both the signal (section header) and the following signal-guided summary.

4.4 Evaluation Methods

Despite recurrent efforts in summarization evaluation, automatic evaluation of generated summaries still has several limitations and biases (Hardy et al., 2019; Fabbri et al., 2021; Ben Abacha et al., 2021). These limitations can misinform current and future research efforts and orient neural networks towards optima that do not accurately reflect the relevance and quality of generated summaries. For instance, commonly used evaluation metrics such as ROUGE do not assess whether summaries are factually consistent with source documents, include critical errors, or lack important information (Goodrich et al., 2019). Manual evaluation is another method to assess the quality of the generated summaries, but is time consuming and relies on the availability of domain experts to rate the summaries. Howcroft et al. (2020) examined 165 NLG papers with human evaluations and concluded that the field is in urgent need of standard evaluation methods and terminology.

Taking into account these different factors, we evaluate the generated summaries using a variety of evaluation metrics based on n-gram matching, pre-trained contextualized embeddings (BERTScore), learned metric (BLEURT), automatic fact-based metrics (Fact Scores), and manual evaluation performed by medical experts. BERTScore (Zhang* et al., 2020) uses the pre-trained contextual embeddings from BERT and matches words in candidate and reference texts by cosine similarity. We use two variants: BERTScore-M1 based on the default roberta-large model and BERTScore-M2 based on the deberta-xlarge-mnli model, which was shown to have the best correlation with human evaluation. BLEURT (Sellam et al., 2020) is a learned metric, based on BERT and a pre-training scheme that uses millions of synthetic examples. We use the latest checkpoint BLEURT-20 that correlates better with human judgment and the F1 variants of all the automatic metrics.

For the fact-based evaluation, we utilize a machine learning-based medical fact extraction system to extract medically relevant facts. A medical fact consists of one core attribute with or without one or more other attributes, such as laterality or bodysite. For example, the medical facts identified by the fact extraction system for the input "*the patient has rash on the upper arms*" are *<FINDING_CORE> rash <LATERALITY> upper <BODYSITE> arms*. The Fact Score metric provides F1-score of the extraction of medically relevant facts. The first variant Fact-Core relies on the extraction of seven core fact attributes: *'Procedure_Core', 'Disorder_Core', 'Finding_Core', 'Medication_Core', 'Substance_Use_Core', 'Vital_Sign_Core', and 'Allergy_Core'*. The Fact-Full variant combines these core facts and five additional attributes: *'Negation', 'Hedge', 'Status', 'Laterality', and 'Bodysite'*.

We also investigate the correlation between the automatic metrics and the expert judgments.

5 Experiments

We train the summarization models on 4 Nvidia Tesla K80 GPUs for 4 epochs. We set the learning rate to 3e-05 and regularize the training with an L2 weight decay of 0.1. We use a test set of 100 conversations and notes, randomly selected from the MTS-DIALOG dataset. The remaining pairs are used for training (1,201 pairs) and validation (400 pairs).

5.1 Automatic Evaluation

Table 6 presents the results of different summarization models fine-tuned and evaluated on the MTS-DIALOG dataset. We picked the best performing model from each category (four models in total) for subsequent studies. Table 7 compares the results of these four summarization models using ROUGE-N, Fact Scores, BERTScore, and BLEURT. Examples of generated summaries are shown in Table 9.

These first results highlight the importance of relevant pre-finetuning targets, with XSum yielding substantially better results as a first pre-finetuning stage compared to CNN/DailyMail (40.15 vs. 32.01 ROUGE-1 score). A portion of this differential could be explained by the relatively short length of the MTS-DIALOG summaries (three sentences and 48 words on average) which is likely closer to the length of output for extreme summarization than the longer output summaries in

	ROUGE-1	ROUGE-2	ROUGE-L
Baseline Models			
Pegasus-large	27.62	10.99	23.03
BART-large [Model#1]	**30.42**	**12.03**	**26.91**
Pre-Finetuning (PFT)			
Pegasus-pubmed	24.20	8.49	18.03
Pegasus-xsum	32.88	13.75	27.43
BART-cnn-samsum	32.01	13.93	23.05
BART-xsum-samsum [Model#2]	**40.15**	**18.04**	**32.56**
Guided Summarization (GS)			
BART-cnn-samsum (GS)	32.68	14.14	23.39
BART-xsum-samsum *(GS)* [Model#3]	**42.04**	**17.59**	**34.85**
GS + Data Augmentation (DA)			
BART-cnn-samsum *(GS+DA)*	33.29	14.58	24.30
BART-xsum-samsum *(GS+DA)* [Model#4]	**42.52**	**17.50**	**34.90**

Table 6: Results of the summarization models fine-tuned and evaluated on the MTS-DIALOG dataset.

	Model#1 (BART-large)	Model#2 (BART-PFT)	Model#3 (BART-PFT-GS)	Model#4 (BART-PFT-GS-DA)
ROUGE-1	0.3042	0.4015	0.4204	**0.4252**
ROUGE-2	0.1203	**0.1804**	0.1759	0.1750
ROUGE-L	0.2691	0.3256	0.3485	**0.3490**
Fact-Core	0.2381	**0.3753**	0.3466	0.3496
Fact-Full	0.1643	**0.2264**	0.2126	0.2128
BERTScore-M1	0.3090	0.3830	**0.4190**	0.4120
BERTScore-M2	0.2850	0.3680	0.4000	**0.4080**
BLEURT-20	0.4316	0.5003	0.5089	**0.5123**

Table 7: Evaluation of the summarization models using lexical, fact-based, embedding-based, and learned metrics (Macro Average F1 scores over the summaries).

CNN/DailyMail.

Data Augmentation (DA) led to slight improvements across all metrics except ROUGE-2 and BERTScore-M1 as shown in Table 7. Guided Summarization (GS) led to a consistent improvement across all automated metrics except for ROUGE-2 and the Fact-based metrics.

5.2 Expert-based Manual Evaluation

The manual evaluation of the generated summaries is performed using NLG criteria such as Fluency and Non-redundancy, and medical criteria such as Critical Omissions based on fact extraction. For this evaluation, we define a fact as information that cannot be written in more than one sentence. For instance: the sentence "*The father died of stroke at age 89.*" could be written in three sentences: "*The father died*", "*He was 89 yo.*", and "*Stroke was the cause of death.*" and thus contains three facts.

The manual summary evaluation criteria are:

- Fluency: Is the summary fluent to read? (0:"none", 1:"low", 2:"average", 3:"high")
- Non-redundancy: How redundant is the summary? (0-3)
- Critical Omissions: What is the number of medical facts that were omitted?
- Hallucinations: What is the number of hallucinated facts?
- Correct Facts: How many facts in the summary are correct according to the input conversation?
- Incorrect Facts: How many facts are incorrect outside of hallucinations (e.g. wrong age)?

We compute the following scores from the manual counts:

$$FactualPrecision = \frac{\#CorrectFacts}{\#SystemOutputFacts}$$

$$FactualRecall = \frac{\#CorrectFacts}{\#ReferenceFacts}$$

$$HallucinationRate = \frac{\#HallucinatedFacts}{\#SystemOutputFacts}$$

$$OmissionRate = \frac{\#OmittedFacts}{\#ReferenceFacts}$$

- *SystemOutputFacts=Correct+Incorrect+Hallucinated*

	Kappa	F1 (exact)	F1 (tol=1)	F1 (tol=2)	Pearson's correlation
Number of key/reference facts	0.494	0.540	0.780	0.930	0.980
Number of correct facts	0.599	0.710	0.930	1.000	0.841
Number of omitted facts	0.366	0.440	0.740	0.890	0.957
Number of incorrect facts	0.305	0.900	0.990	1.000	0.717
Number of hallucinated facts	0.541	0.920	1.000	1.000	0.695
MACRO-AVG	0.467	0.675	0.875	0.957	0.862

Table 8: Fact count agreements.

	Generated Summary	Reference Summary
1	Family history is significant for coronary artery disease, hypertension, diabetes mellitus, and cerebrovascular disease.	Family history is remarkable for heart disease, cerebrovascular disease, diabetes, and hypertension.
2	The patient denies any history of depression, suicidal ideation, chest pain, shortness of breath, nausea, vomiting, numbness, weakness, or tingling.	Please see history of present illness. Psychiatric: She has had some suicidal thoughts, but no plans. She denies being suicidal at the current time. Cardiopulmonary: She has not had any chest pain or shortness of breath. GI: Denies any nausea or vomiting. Neurological: No numbness, weakness or tingling.
3	Significant for frequent flyer status, anemia, anxiety, bipolar disorder, and iron deficiency anemia. Surgery history is positive for tubal ligation.	1. Bipolar disorder. 2. Iron deficiency anemia. 3. Anxiety disorder. 4. History of tubal ligation.
4	He is a non-cigarette smoker and non-ETOH user. He is single and he has no children. He works as a payroll representative and previously did lot of work in jewelry business, working he states with chemical.	She is a nonsmoker and nondrinker. She is single with no children. Currently works as a payroll representative. Prior to that, she worked in the jewellery business with chemical.
5	The patient was a baby born at 32 weeks' gestation at 4 pounds 11 ounces and placed in an incubator for 3 weeks. He had jaundice, but was not given any treatment.	32 weeks gestation to a G4 mother and weighed 4#11oz. He was placed in an incubator for 3 weeks. He was jaundiced, but there was no report that he required treatment.

Table 9: Examples of generated summaries by Model #4.

To assess the effort level for the end user (clinicians who will use/edit the system generated summary), we compute Levenshtein edit distance (minimum # of character insertion, deletion, substitution or transposition operations) (i) between system summary and reference summary, and (ii) between the initial system summary and a human-corrected version of that summary which fixes all issues. The normalized edit distance is then calculated by dividing the Levenshtein distance by the length of the longest summary between reference and system.

To measure the consistency of these expert evaluations, a common set of 100 reference-system outputs were labeled independently by two trained annotators with medical backgrounds. The remainder of the evaluated data (300 system summaries) was single annotated.

The Pearson correlations between the annotations for the rating-based scores were 0.631 for Fluency and 0.894 for Non-redundancy. We also calculated Cohen's kappa and F1 score for the number of reference facts as well as the correct, omitted, hallucinated, and incorrect system facts. With strict Cohen's kappa and F1, we obtained values 0.467 and 0.675 respectively. As these measures penalize harshly for being off by one or two facts, we also measured a relaxed F1 allowing a count mismatch of one or two facts, which showed an inter-annotator F1 agreement of 0.875 and 0.957 respectively (cf. Table 8). The Pearson correlation between the two annotators' fact counts was also high at a macro-average of 0.862.

The results of the manual evaluation are reported in Table 10. In large part they confirm the automatic evaluation results from Table 7, with Models 2-4 performing better than the baseline Model #1 in terms of Factual Recall and Factual F1. In comparison with Model #3, Data Augmentation (Model #4) led to better fluency, non-redundancy, and factual Recall and F1 with less critical fact omissions, but increased the hallucination rate to 3% from 1% for Model #3.

The model without guided summarization or data augmentation (Model #2) achieved similar results to the best model that employed both (Model #4). Model #2 was also ranked higher according to the automatic fact extraction metrics (cf. Table 7). These results suggest that guided summarization improves the precision of the summary facts (+5.5% improvement) but lowers recall (-5%), with data augmentation reversing the trend. Fact-based performance was thus shown to be more

Evaluation Criteria	Model #1 (BART-large)	Model #2 (BART-PFT)	Model #3 (BART-PFT-GS)	Model #4 (BART-PFT-GS-DA)
Summary Quality Evaluation				
Fluency [0-3] ↑	2.33	**2.44**	2.31	2.37
Non-redundancy [0-3] ↑	**2.97**	2.85	2.90	2.93
Fact-based Evaluation				
Factual Precision ↑	**0.9492**	0.8917	0.9408	0.9010
Factual Recall ↑	0.5324	0.6671	0.6341	**0.6685**
Factual F1 Score ↑	0.6822	0.7632	0.7576	**0.7675**
Hallucination Rate ↓	0.02	0.04	**0.01**	0.03
Omission Rate ↓	0.47	**0.33**	0.37	**0.33**
Effort Level Assessment				
Levenshtein Edit Distance [wrt Correction] ↓	0.5770	**0.4944**	0.5858	0.5521
Levenshtein Edit Distance [wrt Reference] ↓	*0.8685*	*0.8426*	*0.8124*	***0.8101***

Table 10: Expert-based Manual Evaluation.

	ROUGE-1	ROUGE-2	ROUGE-L	Fact-Core	Fact-Full	BERT-M1	BERT-M2	BLEURT
ROUGE-1	1.000							
ROUGE-2	0.636	1.000						
ROUGE-L	0.949	0.646	1.000					
Fact-Core	0.410	0.429	0.334	1.000				
Fact-Full	0.379	0.389	0.335	0.740	1.000			
BERT-M1	0.790	0.457	0.801	0.254	0.265	1.000		
BERT-M2	0.857	0.505	0.847	0.362	0.339	0.905	1.000	
BLEURT	0.790	0.429	0.787	0.269	0.247	0.784	0.859	1.000

Table 11: Pearson's correlation coefficients between the automatic evaluation metrics.

sensitive to prefixes in the training data (in the GS experiment we used 20 different section headers as prefixes in the training data). This has likely prevented the models from generalizing factual patterns across different sections, even though the same prefixes helped improve the performance on a token level according to the token-based evaluation measures (ROUGE-1 and BERTScore). On the other hand, guided summarization improved non-redundancy and lowered the factual hallucination rate to only 1% down from 4% for Model #2.

Our four models provided summaries with an average length of 9.76, 24.6, 19.77, and 21.45 tokens, respectively. To evaluate potential performance bias from summary length, we computed the correlation between the summary length and BLEURT as an example of automatic metric and the correlation between the summary length and Factual F1 as an example of manual metric. The Pearson correlation scores between BLEURT and the summary length of Model #1, #2, #3, and #4 are low (-0.200, -0.027, -0.147, and -0.173, respectively). The correlation between Manual Factual F1 an the summary length of the same models are higher with an inverse correlation of -0.511, -0.230, -0.268, and -0.409, respectively, which indicates that the models are prone to more errors in longer summaries.

5.3 Correlation between Evaluation Metrics

Table 11 shows Pearson correlations between the automatic evaluation metrics. BERTScore-M2, based on the DeBERTa model, was the embedding-based metric with the highest correlation with the n-gram metrics ROUGE-1, ROUGE-2, and ROUGE-L, and the embedding-based metrics BERTScore-M1 and BLEURT. However, ROUGE-1 and ROUGE-2 had higher correlation with Fact-Core and Fact-Full.

Table 12 presents the Pearson correlations between the automatic metrics and manual scores. The correlations with the manual scores show a different picture, with BLEURT standing out as the most correlated with manual fact counts and expert-based correctness assessments. BLEURT was also the most correlated metric with the Levenshtein distance, used here as an indicator for the level of effort required to correct the generated summaries.

The manual fact metrics were more correlated with the embedding-based metrics than ROUGE-1, ROUGE-2, and ROUGE-L, in contrast with the automatic fact extraction metrics. This could be explained, in part, by the lower factual coverage of automatic fact extraction vs. manual fact identification. This also highlights an important empirical insight, in that upstream upper bounds, such as the limited coverage of automatic and symbolic fact

Automatic \ Manual	Factual P	Factual R	Factual F1	Hallucination	Omission	Levenshtein
		Correctness ↑			Error Rate ↓	
ROUGE-1	0.101	0.368	0.402	-0.074	-0.486	-0.326
ROUGE-2	0.086	0.167	0.202	-0.049	-0.237	-0.040
ROUGE-L	0.126	0.393	0.416	-0.073	-0.479	-0.296
Fact-Core	0.039	0.095	0.160	**-0.105**	-0.192	-0.030
Fact-Full	0.056	0.133	0.188	-0.094	-0.214	-0.122
BERTScore-M1	**0.132**	0.445	0.462	-0.078	-0.518	-0.317
BERTScore-M2	0.090	0.437	0.461	-0.080	-0.562	-0.366
BLEURT-20	0.113	**0.477**	**0.480**	-0.082	**-0.591**	**-0.486**

Table 12: Pearson's correlation coefficients between the automatic and manual scores.

extraction, can bias the correlation results against neural embeddings methods which, from their large pre-training, have inherently wider (implicit) coverage than symbolic fact extraction methods.

6 Conclusion

In this paper we conducted an empirical study of clinical note generation from doctor-patient encounters. This included creating a new dataset of 1,700 conversations and notes, evaluating several SOTA summarization models, and using multiple automated metrics and human judgements for the summarization models. Our findings show that pre-finetuning transformer models plays a key role in improving factual accuracy and summary fluency, and in reducing critical fact omissions, with guided summarization improving the precision of the summary facts and reducing hallucinations at the expense of factual recall.

The manual evaluation showed that the generated summaries reached a high fluency score (2.44 on a 0-3 scale) and relatively high factual F1 (0.76), but the best model still had a hallucination rate of 3% and missed 33% of the medical facts. Wider research efforts are needed to design methods and models that can reduce the omission and hallucination rates as well as efficient fact-based evaluation metrics to automatically assess the factual consistency of the generated summaries.

The key bottleneck limiting increased research from the NLP community on medical note generation has been the lack of generally available datasets to train and experiment on. This public release of the MTS-DIALOG dataset will enable wider research and faster research progress on this very important and impactful NLP task. Our comprehensive evaluation of multiple summarization models and evaluation metrics provides valuable baselines and references for comparison, and also shows where these systems and metrics stand on multiple dimensions of human evaluation.

Medical doctors today suffer from heavy documentation burden, which frequently causes physician dissatisfaction and burnout, and distracts doctors from being able to give their undivided attention to their patients. We look forward to a future where NLP research and the NLP community are able to provide doctors with tools for effective automated medical note generation, and enable them to return their focus to what they really love - providing great care for their patients.

Limitations

To address the lack of data and protect patient privacy, we relied on creating simulated doctor-patient conversations from de-identified clinical notes. Although we relied on trained annotators with medical background, those simulated conversations might still not reflect faithfully the actual language or structure of real doctor-patient conversations. For instance, the fraction of disfluencies and speech interruptions may occur at higher frequency than simulated in this dataset. Furthermore, it is often the case that doctor-patient conversations are automatically transcribed from speech to text, and the speech-to-text generation systems can produce errors in the output text. If real conversation benchmarks are made available in the future, they would allow further extrinsic validation of the proposed dataset and the first research insights. Many more summarization models can be tested on the task with the new dataset, and this study explored only some of them. Different, or larger language models could yield different results and lead to new or different insights. Although the introduced dataset is substantially bigger in size compared to other doctor-patient conversation datasets, its size is still limited in comparison with open-domain summarization datasets, which can limit the fine-tuning performance, especially for larger neural

models. The dataset is limited in terms of covered diseases. The annotators ensured that the dataset is fully anonymized and balanced the synthetic conversations w.r.t. gender but the notes could still include real-world biases from the original clinical notes. The created dataset is intended for research purposes on automatic clinical note generation and should not be used for medical diagnosis or other health-related applications.

Ethics Statement

No protected health information were used in the creation of this dataset. Annotators were paid a fair hourly wage consistent with the practice of the state of hire.

References

Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. 2017. Tethered to the EHR: Primary care physician workload assessment using EHR event log data and time-motion observations. In *Annals of Family Medicine*, volume 15(5), pages 419–426.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis P. Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP@NAACL-HLT 2021, Online, June 11, 2021*, pages 74–85. Association for Computational Linguistics.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41:181–190.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *CoRR*, abs/1911.12237.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*, pages 735–742. IEEE.

Ben Goodrich, Vinay Rao, Mohammad Saleh, and Peter J. Liu. 2019. Assessing the factual accuracy of generated text. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Hardy, Shashi Narayan, and Andreas Vlachos. 2019. Highres: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3381–3392. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 169–182. Association for Computational Linguistics.

George Hripcsak, David K Vawdrey, Matthew R Fred, and Susan B Bostwick. 2011. Use of electronic clinical documentation: time spent and team interactions. *Journal of the American Medical Informatics Association*, 18:112–7.

Adam L. Janin, Don Baron, Jane Edwards, Daniel P. W. Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The icsi meeting corpus. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, 1:I–I.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3755–3763. Association for Computational Linguistics.

Nazmul Kazi and Indika Kahanda. 2019. Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations. In *Proceedings of the 2nd Clinical Natural Language Processing*

Workshop, pages 140–148, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos-Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. *CoRR*, abs/2205.02549.

Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4958–4972. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2190–2196. Association for Computational Linguistics.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5032–5041. Association for Computational Linguistics.

Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhuguo Li, and Xiaodan Liang. 2020. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *ArXiv*, abs/2010.07497.

Zhengyuan Liu, A. Ng, Sheldon Lee Shao Guang, AiTi Aw, and Nancy F. Chen. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, volume EMNLP 2022 of *Findings of ACL*, page 4741–4749. Association for Computational Linguistics.

Francesco Moramarco, Damir Juric, Aleksandar Savkov, Jack Flann, Maria Lehl, Kristian Boda, Tessa Grafen, Vitalii Zhelezniak, Sunir Gohil, Alex Papadopoulos Korfiatis, and Nils Hammerla. 2022a. Towards more patient friendly clinical notes through language models and ontologies. In *AMIA Annu Symp Proc*.

Francesco Moramarco, Alex Papadopoulos-Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022b. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5739–5754. Association for Computational Linguistics.

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. *CoRR*, abs/1808.08745.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Jamin Shin, Hangyeol Yu, Hyeongdon Moon, Andrea Madotto, and Juneyoung Park. 2022. Dialogue summaries as dialogue states (ds2), template-guided summarization for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3824–3846. Association for Computational Linguistics.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Brett R. South, Danielle L. Mowery, Ying Suo, Jianwei Leng, Óscar Ferrández, Stéphane M. Meystre, and Wendy W. Chapman. 2014. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. *J. Biomed. Informatics*, 50:162–172.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Wen-wai Yim and Meliha Yetisgen. 2021. Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3693–3712. Association for Computational Linguistics.

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

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. MIE: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469, Online. Association for Computational Linguistics.