Generation of Patient After-Visit Summaries to Support Physicians

Pengshan Cai,[1] **Fei Liu,**[2] **Adarsha Bajracharya,**[3] **Joe Sills,**[3] **Alok Kapoor,**[3]
Weisong Liu,[4] **Dan Berlowitz,**[4] **David Levy,**[4] **Richeek Pradhan,**[5] **Hong Yu**[1,3,4]
[1]University of Massachusetts, Amherst [2]Emory University
[3]UMass Chan Medical School [4]University of Massachusetts, Lowell [5]McGill University
{pengshancai, hongyu}@cs.umass.edu fei.liu@emory.edu
{weisong_liu, dan_berlowitz, david_levy}@uml.edu
{adarsha.Bajracharya, alok.kapoor}@umassmemorial.org
joe.sills@baystatehealth.org richeekp@gmail.com

Abstract

An after-visit summary (AVS) is a summary note given to patients after their clinical visit. It recaps what happened during their clinical visit and guides patients' disease self-management. Studies have shown that a majority of patients found after-visit summaries useful. However, many physicians face excessive workloads and do not have time to write clear and informative summaries. In this paper, we study the problem of automatic generation of after-visit summaries and examine whether those summaries can convey the gist of clinical visits. We report our findings on a new clinical dataset that contains a large number of electronic health record (EHR) notes and their associated summaries. Our results suggest that generation of lay language after-visit summaries remains a challenging task. Crucially, we introduce a feedback mechanism that alerts physicians when an automatic summary fails to capture the important details of the clinical notes or when it contains hallucinated facts that are potentially detrimental to the summary quality. Automatic and human evaluation demonstrates the effectiveness of our approach in providing writing feedback and supporting physicians.[1]

1 Introduction

Studies have shown that the majority of patients do not understand their clinical visits (O'Leary et al., 2010). After-visit summary note (AVS) is a summary given to patients after their clinical visit, it is intended to summarize patients' clinical visits and help their disease self-management (Federman et al., 2018). Compared to clinical notes, an after-visit summary has the following characteristics: 1) it is written in lay person language thus is easy for patients to read and comprehend; 2) it only contains information that patients should be aware of, leaving out redundant details, e.g. unimportant lab results, etc. Studies have shown that around 36% of American adults have limited health literacy (Kutner et al., 2006), and 94.4% of patients found that lay language after-visit summary helps them understand their clinical visits (Pathak et al., 2020). However, the implementation of after-visit summary is challenging. Many physicians face excessive workloads (West et al., 2018) and do not have time to complete the summaries in a timely manner (Hong et al., 2013). Thus, there is a real need for—and this study contributes to—automatic generation of after-visit summaries to unburdening physicians with complex information workflows.

We explore best-performing neural abstractive summarizers to generate after-visit summaries from EHR notes. The summaries are rated by physicians as concise and easy to read. However, they can not be presented directly to patients, as they frequently contain two types of errors: 1) *Missing content.* A summary often leaves out important details such as medication dosage and route, undermining patients' medical self-management. 2) *Hallucination.* Summaries contain hallucinated content or content not supported by the input documents. For example, an abstractive summary on *kidney infection* was generated from an input document that describes *urine infection*. These types of errors are not uncommon in abstractive summarization (Lebanoff et al., 2019; Maynez et al., 2020; Pagnoni et al., 2021), but they could be disastrous to patients.

In this study, we build systems to facilitate detection and correction of those types of errors, allowing physicians to correct or edit system generated summaries. As illustrated in Figure 1, *Summarization* produces a system summary; *Error Alerting* automatically detects errors from the generated after-visit summary. Crucially, we build effective detectors with self-supervision on unlabeled data for error alerting. A novel dataset is constructed by synthesizing summaries containing medical events that are inconsistent with their source documents.

[1]Our project page is available at: https://github.com/pengshancai/AVS_gen

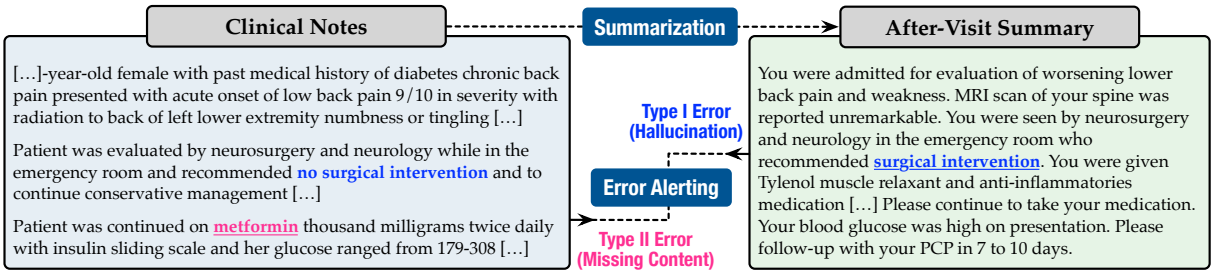

Figure 1: An example after-visit summary generated from EHR notes associated with a patient. A novel alerting mechanism is proposed in this work to report errors found in the summary, including missing medical events and hallucinated facts. We aim to build effective detectors with self-supervision on unlabeled data for error alerting.

Using this simulated dataset, we train a hallucination detection model, which alerts physicians of potential hallucination content. Further, by aligning medical events in EHR notes to those in after-visit summaries using MetaMap (Aronson, 2001), we identify key events important to patients, and alert physicians of salient medical events not covered in the generated summaries as missing content. The contributions of our research are as follows:

- We propose a new task that generates lay language AVS from EHR notes, build and evaluate state-of-the-art NLP models for this task. A novel alerting mechanism is proposed to report errors, including missing medical events and hallucinations. The training of our error detectors is self-supervised, using only unlabelled text.
- Clinical applications demand high performance. Existing automatic metrics are not adequate for evaluating the quality of generated AVS. Therefore, we conduct a qualitative assessment of system outputs with medical practitioners. Our findings show that the alerting mechanism could provide a promising avenue towards making the writing process easier for physicians.

2 Related Work

Recently there has been a lot of work on automatic summarization in the clinical domain: Zhang et al. (2018) propose to generate the impression section of a radiology report using seq2seq models. Miura et al. (2021) perform image-to-text radiology report generation by optimizing entity-based rewards with reinforcement learning. Studies are also performed for summarizing doctor-patient dialogues (Joshi et al., 2020; Krishna et al., 2021) and evaluating system generated notes (Moramarco et al., 2022).

Early work has explored generation of hospital visit summaries using non-neural methods (Di Eugenio et al., 2014; Hirsch et al., 2015; Acharya et al., 2018). Recently, Adams et al. (2021) present the task of hospital-course summarization with the goal of generating a text to synthesize the hospital course. A crucial difference between our work and that of Adams et al. (2021) is we investigate a deep learning solution, whose primary focus is exposing neural abstractive summarizers to clinical notes and explicitly highlighting regions of a summary which need attention. This is in principle similar to Checklist in (Ribeiro et al., 2020).

It is important for an after-visit summary generated from EHR notes to avoid type I and type II errors. A type I error (false positive) suggests that there is false or inaccurate information in the summary, due to hallucinations, incorrect grounding, *etc*. It is a challenging and lingering problem facing natural language generation (Falke et al., 2019; Lebanoff et al., 2020; Kryscinski et al., 2020; Matsumaru et al., 2020; Pagnoni et al., 2021; van Miltenburg et al., 2021), despite remarkable recent progress (Gehrmann et al., 2018; Liu and Lapata, 2019; Fabbri et al., 2019; Zhong et al., 2020; Lewis et al., 2020; Ni et al., 2021, *inter alia*).

A more surprising observation is that the type II error (false negative) is deemed particularly harmful to patients. When salient medical events such as diagnoses or treatments are left out of the after-visit summaries, it could have a detrimental effect on patients' self-care after being discharged from hospitals (Raghavan et al., 2012; Sotudeh Gharebagh et al., 2020). This empirically motivates our work, where we seek to effectively identify salient medical events in EHR notes and alert physicians of any missing events to help them avoid those errors.

A distinguishing characteristic of after-visit summaries is that they are *patient-oriented*. The summaries provide relevant and actionable information to patients, such as reasons for visit, diagnoses and procedures, *etc*. Differing from *physician-oriented*

Abstractive Summarization Model	Extractive Summarization Model
• **BART** (Lewis et al., 2020) uses the standard encoder-decoder architecture. It was pretrained as a denoising autoencoder to learn to reconstruct the original text. Our input to the BART model consists of a clinical document and its output is an abstractive summary. • **PEGASUS** (Zhang et al., 2020a) explores a new pretraining objective tailored for abstractive summarization. Important sentences are masked out from the input document and the model learns to generate the sentences as an output sequence, akin to an extractive model. The system has been shown to perform well in a low-resource scenario where few examples are available for fine-tuning. • **LED** (Beltagy et al., 2020) is the Longformer-Encoder-Decoder model. It is an extension to Longformer to support text generation. LED uses a local windowed attention which makes it computational feasible to encode a long input document. We favor the LED model because, compared to news articles, there is more risk involved in truncating long clinical documents to a certain length.	• **BertSum** (Liu and Lapata, 2019) employs the BERT model to identify summary-worthy sentences. It uses a flat architecture to encode the input document, then adds a Transformer layer on top of the sentence representations to model inter-sentence relationship. The final output layer is a sigmoid classifier used to predict if the sentence is to be included in the summary. • **TextRank** (Mihalcea and Tarau, 2004) and **LexRank** (Erkan and Radev, 2004) are graph-based models that extract relevant sentences based on eigenvector centrality. • **Oracle Top-K** (Adams et al., 2021) is a method introduced by Adams et al. which represents the upper bound for sentence extraction. It ranks all document sentences according to their averaged R-1 and R-2 scores with respect to the reference summary. It then continues to add sentences yielding the highest scores to the summary until the target token count is reached.

Table 1: State-of-the-art summarization models investigated in this work for generation of patient after-visit summaries.

clinical notes, these summaries are written in an easy-to-understand language, and they remain understudied in NLP. Existing research on text simplification focuses primarily on Wikipedia and news articles (Zhu et al., 2010; Xu et al., 2015; Vu et al., 2018; Kriz et al., 2019; Dong et al., 2019; Kriz et al., 2020). Chandrasekaran et al. (2020) propose to generate lay summaries to describe scientific papers for non-experts. In a similar fashion, after-visit summaries are intended for a lay audience: translating sophisticated medical events into plain language that is understandable by patients.

In what follows, we investigate generation of patient after-visit summaries, closely examine the language used by clinicians, and develop automatic methods to spot errors in summaries to better support clinicians with this challenging task.

3 Summarization

Our method generates an after-visit summary from EHR notes concerning a patient. It is modelled as a single-document summarization task as the EHR notes were collapsed into a single document by the hospital and we were unable to recover individual EHR notes. We use $\mathcal{S}$=$\{w_1, ..., w_{n_\mathcal{S}}\}$ to denote tokens of the source document and $\mathcal{T}$=$\{w_1, ..., w_{n_\mathcal{T}}\}$ tokens of the target summary, $n_\mathcal{S}$ and $n_\mathcal{T}$ are length of the sequences.

We explore a variety of summarization models to generate after-visit summaries. They are detailed in Table 1. Particularly, an abstractive summarizer employs the standard Transformer-based encoder-decoder model to generate a summary $P(\mathcal{T}|\mathcal{S})$. An extractive summarizer selects important sentences to add to the summary until a length threshold has been reached. These systems are used off-the-shelf and have achieved some of the highest reported scores on summarization. We assess their ability to navigate complex medical terrain for generation of after-visit summaries.

Entity Type	Event Type
Anatomical Abnormality	Diagnostic Procedure
Medical Device	Therapeutic or Preventive Procedure
Clinical Drug	Pathologic Function
Pharmacologic Substance	Disease or Syndrome
Organic Chemical	Mental or Behavioral Dysfunction
Body Substance	Injury or Poisoning
Finding	
Sign or Symptom	

Table 2: Semantic types used in this study.

Clinical notes are complex and full of references to medical events. However, the summary given to the patient is simple and clear. We are thus curious to know how medical events manifest themselves in the context of summarization. Events are especially important for this task, as salient events happening at each medical encounter must be included in the after-visit summary.

We define *event nugget* as a word or multi-word phrase that clearly expresses the occurrence of a medical event. Event nuggets are identified by MetaMap (Aronson, 2001), an open-source software tool designed to discover medical concepts referred to in a text. Each occurrence of the concept is assigned a concept unique identifier (CUI) and its associated words are tagged in the text. In Figure 3, we show an example of medical concepts identified by MetaMap. Further, those medical concepts are categorized into various semantic types. We focus on concepts pertaining to a selected set of entity and event types (Table 2), which are deemed relevant by medical experts. The other types are excluded from consideration.

4 Error Detection and Alerting

Event nuggets are associated with type I and type II errors frequently found in the summary. A *type I error* indicates a summary contains a hallucinated fact or event that is not present in the source document. A *type II error* suggests that an important medical event has been mistakenly left out of the summary, hampering its usability. In this section, we describe novel methods to detect likely errors and flag them in the text to alert clinicians.

4.1 Type I Error: Hallucination

A *hallucination detector* aims to recognize any hallucinated content in a summary. Flagging errors is helpful because physicians can be alerted about any anomalies and it is especially appreciated in medical domain (Singh et al., 2014). Our detector uses the BigBird model (Zaheer et al., 2020), which is an encoder-only architecture equipped with sparse attention to reduce Transformer's quadratic complexity to linear, and capable of encoding thousands of tokens. The model takes as input a source document ($\mathcal{S}$) and its system summary ($\mathcal{T}$), and outputs a sequence of binary labels, one for each summary token, where 1 represents the token is considered hallucinated and 0 otherwise.

A key factor to the success of our model is its self-supervised training, where a large number of training instances are constructed from unlabeled data. Each training instance is a *synthesized summary* whose hallucinated tokens are flagged. We adapt the model of Zhou et al. (2021), initially proposed for MT, to create our training instances. Our method differs from theirs in that, synthesized summaries are required to contain hallucinated medical events that are inconsistent with or unjustified by the source document.

Synthesizing Erroneous Summaries. The procedure for generating synthesized summaries is illustrated in Figure 2. Given a summary sentence, we mask out one or two of its event nuggets. It is then fed to a denoising auto-encoder (Lewis et al., 2020) to produce an output sentence, whose masked-out positions are refilled with medical events that are "hallucinated" by the model. If the output is substantially different from the input, e.g., with <50% token overlap, it is called a *synthesized* sentence with hallucinations. Tokens of the synthesized sentence, which cannot be aligned to the original sentence using an edit-distance-style algorithm, are flagged. E.g., "*cardiac catheterization*" and "*abnormalities*"

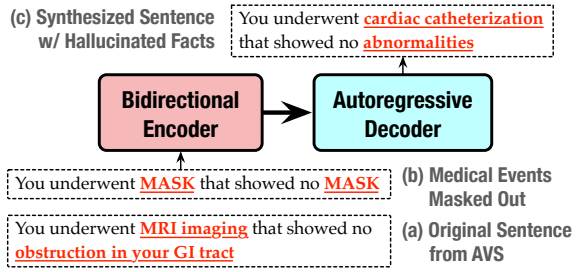

Figure 2: One or two event nuggets are randomly masked out from a summary sentence (a). The masked sequence (b) is fed to a denoising auto-encoder to produce a synthesized sentence that may contain hallucinated medical events (c).

in our example are clearly hallucinated facts. This procedure is repeated for all sentences[2] of the reference summary to create a synthesized summary. We provide examples of synthesized sentences in the Supplementary.

Importantly, the model is fine-tuned to enable it to produce plausible synthesized summaries. We partition the training data into K folds (K=5) of roughly equal size. The BART model is fine-tuned on the union of the K-1 folds, then applied to the remaining fold to generate synthesized summaries. The method transforms each reference summary of the dataset to a synthesize summary, which together with the source document, is used to train and test our type I error detector.

4.2 Type II Error: Missing Content

Our *missing content detector* seeks to accomplish two objectives: 1) to detect *salient medical events* on a clinical document, and 2) to flag salient events that are *missed* by the summary. It is a non-trivial task to fulfill these objectives. Even though clinical notes are full of references to medical events, only a selective portion of them (≈18%) are included in after-visit summaries. As such, we formulate the problem as a classification task. An *event nugget* is assigned a label of 1 if it is salient, 0 otherwise. Our detector leverages self-supervised learning to identify salient events on EHR notes. It then alerts clinicians if the summary fails to include any of the salient events.

Pseudo-Annotations for Salient Events. We create pseudo-annotations for salient events by aligning each source event with one of the target events. As shown in Figure 3, the medial events are identified by MetaMap (Aronson, 2001). Each

[2]If a summary sentence does not contain any medical event, it is left as-is in the synthesized summary. The original summary sentence is otherwise replaced by a synthesized sentence.

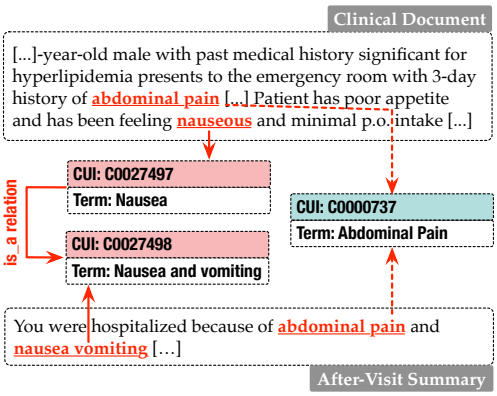

Figure 3: "*abdominal pain*" appears in both the clinical document and after-visit summary, with the same CUI. "*nausea vomiting*" and "*nauseous*" are aligned because there is an *is-a* relation between the two concepts.

occurrence of the event is associated with a concept unique identifier (CUI). Under the *strict matching* criterion, an event of the clinical document is labeled as 1 if an exact match (with the same CUI) is found in the summary. However, a large number of events are not well-aligned under this criterion due to distinct expressions used in clinical notes and summaries. This discrepancy in language use has its origin—clinical notes are physician-oriented, whereas after-visit summaries are patient-oriented. We explore *lenient matching* to alleviate mismatch. If a source event can reach any of the target event via *a single hop* on the UMLS semantic graph,[3] the source event is labeled as salient. In Figure 3, source event "*nauseous*" is leniently matched to target event "*nausea vomiting*," because there exists an "*is_a relation*" between the two events.

We fine-tune the BigBird model (Zaheer et al., 2020) to detect salient events. Being an encoder-only model, BigBird constructs contextualized representations for all tokens of a clinical document. It does not directly produce event representations. To address this issue, we let the model predict salient tokens during training. If a token is part of a salient source event, its gold-standard label is 1. At test time, the model generates token-level predictions. A source event is considered salient if any of its tokens is labeled as 1.

We explore two variants of the model to allow it to better capture events. Both variants aim to inform the model about the occurrences of event nuggets identified by MetaMap. The first variant, +POS, modifies the source sequence by inserting special tokens respectively at the beginning and end

[3] https://www.nlm.nih.gov/research/umls/META3_current_relations.html

BASE +POS	[E] **CT scan** [/E] showed worsening of his [E] **diverticulitis** [/E] with a 5.6 x 3.9cm multiloculated fluid collection in his abdomen.
BASE +TYPE	[Type1] **CT scan** [/Type1] showed worsening of his [Type2] **diverticulitis** [/Type2] with a 5.6 x 3.9cm multiloculated fluid collection in his abdomen.

Table 3: Model variants +POS and +TYPE aim to inform the model about the occurrences of events identified by MetaMap.

of a candidate event. The second variant, +TYPE, inserts different special tokens such that they correspond to the semantic types of the events. We conjecture that certain event types, e.g., *body substance*, are more likely be considered insignificant. In Table 3, we provide examples comparing the source sequences used by model variants.

5 Experiments

In this section, we describe our dataset, perform in-depth analyses on our models, and discuss feedback from physicians who participated in our qualitative evaluation.[4]

5.1 Dataset

Through a collaboration with University of Massachusetts Chan Medical School, we are able to use their electronic health record database, which gives us access to 31,895 EHR notes and their physician-written summaries. All medical records are de-identified to protect patient privacy. These patients were admitted to the medical and surgical services of the hospital from October 2017 to March 2020.

Table 4 summarizes the statistics of our dataset. It is divided into train, validation, and test sets containing 28,157, 1,884 and 1,854 instances, respectively. The dataset has unique characteristics. We observe that the source documents are substantially longer and contain more medical events than their summaries. This is because most hospitalization details are omitted for patients. In addition, the length of clinical documents varies considerably, so is the case for summaries. A long clinical document could be the result of an extended hospital stay. An after-visit summary could be long or short depending on the patient's medical conditions. In contrast, variation in length is less significant in other genres such as news and scientific articles.

We find that an average summary contains 12.3 medical events, yet only 7.9 of them can be linked

[4] Implementation details, including hidden state sizes, computational infrastructure used, hyperparameter configurations, etc. are provided in the Supplementary for reproducibility.

Train / Validation / Test Split		28,157 / 1,884 / 1,854
Number of words per clinical document		523.6 ± 464.3
Number of words per after-visit summary		153.5 ± 166.7
Event Nuggets	per clinical document	42.8 ± 32.0
	per after-visit summary	12.3 ± 9.0
	occurring in both (*lenient match*)	7.9 ± 6.1
	occurring in both (*strict match*)	4.0 ± 3.9

Table 4: Statistics of our dataset.

Adequacy	
3	AVS contains all the information the patient needs to know
2	AVS misses some (1-3) points the patient needs to know
1	AVS misses more than 3 points
Faithfulness	
3	AVS contains no or only a few errors that are ignorable
2	AVS contains some (1-3) factual errors
1	AVS contains more than 3 factual errors
Readability	
3	AVS is easy to read for a lay person
2	AVS has some (1-3) points hard to be understood by the patient
1	AVS has more than 3 points hard to be understood by the patient
Ease of Revision	
3	Physician may spend <=2 minutes to revise the AVS
2	Physician may spend >2 minutes to revise the AVS
1	Physician prefers to not revise the AVS but rewrite from scratch

Table 5: Instructions provided to physicians. The scoring scale for summary evaluation is from 1 (worst) to 3 (best).

to events of the clinical document. The gap is partially due to using MetaMap for medical event identification (Reátegui and Ratté, 2018), which has a reported F-Score of 0.88 and may miss out-of-vocabulary event tokens. Additionally, physicians may add their instructions directly to patient's after-visit summaries, and such content is not grounded in clinical documents.

5.2 Evaluation Metrics

Quantitative Measures. We evaluate the performance of our summarization and error alert models with a variety of quantitative measures.

- **ROUGE** (Lin, 2004) is the standard measure for summarization evaluation. It assigns a high score to a system summary if it has lexical overlap with the reference summary.
- **BERTScore** (Zhang et al., 2020b) is one of the new evaluation metrics for natural language generation that are built on contextualized representations produced by BERT and similar models.
- **SARI** (Xu et al., 2016) is widely used for simplification. It counts how often a system summary correctly keeps, deletes, and adds n-grams.
- **DaleChall** (Dale and Chall, 1948) calculates the readability of the summary based on its sentence length and number of difficult words in it. It is an improvement upon Flesch's reading ease score.
- **P/R/F** scores are reported for error alert models on successful detection of missing medical events and detection of hallucinated summary tokens.

Qualitative Measures. In high-stake scenarios, automatic metrics alone cannot guarantee a good system. Thus, we need expert assessments by medical practitioners in this study. We recruit six human evaluators: five of them are physicians with M.D., one is a M.D. student. Owing to budget constraints, we select a random set of 18 clinical documents and their best system summaries for qualitative assessment. The system summaries are produced by the LED model, they are abstractive. Each summary is judged by two human evaluators, who perform two tasks on a summary:

- **Scoring.** A summary is rated along four dimensions. *Adequacy*: Does the summary contain all necessary information for the patient to know? *Faithfulness*: Does the summary faithfully convey the content of the clinical document? *Readability*: Is the summary easy to read for a lay person? *Ease of Revision*: How long might it take for a physician to revise the summary to meet the expectations of standard AVS? The scoring scale is from 1 (worst) to 3 (best). Their interpretations are provided in Table 5.
- **Revision.** We ask human evaluators to edit the summary until it meets the expectations of standard after-visit summaries. We report the edit distance between the original and edited summaries, the amount of editing applied to the raw system summary is a good indicator of its utility (Snover et al., 2006).

For alert evaluation, we ask the evaluators to first label missing medical events on the clinical document, and hallucinations on the system summary. The evaluators are then given the alerts produced by our models, and they proceed to judging the correctness of each alert. This allows us to report precision, recall and F1 scores of our error alert models with human judgment.

5.3 Summarization Results

Quantitative. Table 6 provides a quantitative evaluation of after-visit summaries produced by state-of-the-art models. Our aim in this work is not to present new methods, but rather to thoroughly evaluate state-of-the-art models on this challenging

	Model	R-1	R-2	R-3	R-4	R-L	BertS	SARI	DaleC.↓	Length
EXT	TEXTRANK (Mihalcea and Tarau, 2004)	25.71	7.36	3.92	2.64	13.83	54.37	34.33	12.56	150.01
	LEXRANK (Erkan and Radev, 2004)	25.57	7.26	4.01	2.71	12.81	54.31	34.91	12.38	153.21
	BERTSUM (Liu and Lapata, 2019)	26.22	7.42	4.43	2.90	14.56	55.62	35.57	11.21	149.73
	ORACLE (Adams et al., 2021)	36.84	13.55	6.86	4.45	19.47	58.50	39.74	11.07	99.61
ABS	BART (Lewis et al., 2020)	41.67	21.05	14.20	10.80	30.20	62.80	44.36	9.97	144.29
	PEGASUS (Zhang et al., 2020a)	37.02	19.68	14.02	10.93	28.44	60.91	41.89	10.53	134.26
	LED (Beltagy et al., 2020)	**41.96**	**21.80**†	**15.01**†	**11.58**†	**31.49**†	**63.31**†	**45.06**	**9.58**	148.03

Table 6: Quantitative evaluation of patient after-visit summaries produced by state-of-the-art summarization models. LED shows best performance among all tested abstractive models. It significantly outperforms all other systems for all metrics (p<0.05), with the exception of BART in terms of R-1, according to a non-parametric Wilcoxon signed rank test.

task to identify areas for improvement. We observe that BERTSUM achieves the highest scores among all extractive models. Further gain is provided by an *oracle* model developed by Adams et al. (2021) that improves R-2 F-score from 7.42% to 13.55% by greedily extracting sentences yielding highest similarity scores with the reference summary. The method gives an upper bound on ROUGE scores obtainable by an extractive model.

We find all abstractive models to perform substantially better than their extractive counterparts. LED has shown best performance among all tested abstractive models, possibly due to its exceptional ability to encode long documents. With regards to evaluation metrics, we include less commonly used R-3 and R-4 F-scores, as they have been shown to correlate better with human judgment than other variants (Graham, 2015; Kryscinski et al., 2019). Our results suggest that generation of patient after-visit summaries is highly abstractive. For this reason, an abstractive model would suit our task best. Extractive summaries are verbose and they may potentially overwhelm patients with unnecessary detail.

Expert Scoring. Two medical experts are asked to rate each summary produced by our best abstractive model (LED) along the dimensions of *adequacy*, *faithfulness*, *readability* and *ease of revision*. Their ratings are averaged for each summary [5] and results are presented in Figure 4. All summaries are divided into five bins, their average ratings are 1/1.5/2/2.5/3, respectively. We observe that generating adequate summaries remains a challenge for the abstractive model. Only 5.5% of the summaries obtain a full score (3 points). Per our physicians, the remaining summaries have, to a varying degree, missed important medical events that patients need to know. Our findings suggest that future studies should incorporate expert knowledge in selecting medical events to add to the summary.

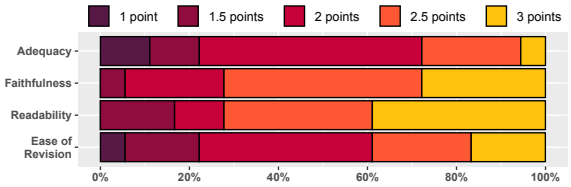

Figure 4: Summaries are rated by medical practitioners along the dimensions of *adequacy*, *faithfulness*, *readability* and *ease of revision*. Their ratings are averaged for each summary.

Efforts could be made to also improve the readability and understandability of abstractive summaries. We observe that 38.8% of the summaries obtain a full score on readability. A closer analysis reveals that a portion of the summaries contain abbreviated medical terminology or jargon that are familiar to physicians but may be difficult for non-experts. E.g., in "*minimal PO intake*," PO is from the Latin "*per os*" and means "*by mouth*." The summaries are also believed to have less hallucination issues when comparing to missing medical events. 72.2% of the summaries obtain 2.5 points or higher. Further, >75% of the summaries receive an average score of 2.5 or higher on ease-of-revision. The results indicate that, physicians may be guided to revise system-produced summaries to meet the standards of medical practice, as opposed to starting from scratch.

Expert Revision. Table 8 shows a direct comparison of summaries before and after expert revision (more examples are in the supplementary). Our physicians have revised 43.5 words on average for each summary, corresponding to 47.2% of the summary length. Even though there is still room for improvement, the results are positive. For 4 out of 18 cases, physicians only minimally revised the summaries, with less than 15% of the

[5] We provide inter-annotator analysis among physicians in the supplementary materials.

	Model	P(%)	R(%)	F_1(%)
Type I	Baseline-RAND	6.52	3.22	3.82
	Baseline-MOSTFREQ5	17.80	46.04	21.79
	Baseline-MOSTFREQ10	20.86	76.31	29.19
	H-Alert (Ours)	**44.96**	**71.66**	**55.25**
Type II	Baseline-RAND	3.99	13.56	6.06
	Baseline-MOSTFREQ5	9.74	34.72	13.70
	Baseline-MOSTFREQ10	9.71	49.14	15.04
	M-Alert (Ours)	49.22	43.65	41.71
	M-Alert +POS (Ours)	**51.03**	45.98	43.80
	M-Alert +TYPE (Ours)	50.69	**49.88**	**45.51**

Table 7: Automatic evaluation of our hallucination detector (**H-Alert**) and missing event detector (**M-Alert**). Both detectors strongly outperform their baselines.

words edited. For 3 out of 18 cases, the summaries are nearly rewritten, where 90% of the words are edited. The results suggest that certain noisy clinical documents can cause disastrous summaries. It is crucial for summarizers to degrade gracefully as noise increases.

5.4 Error Detection Results

Our detectors are evaluated using both automatic metrics and human judgment. Results are reported in Table 7. **H-Alert** is our hallucination detector. It is evaluated on the test set with synthesized hallucinations (§4.1). Baseline-RAND samples a label for each summary token from a Bernoulli distribution $t_j \sim \text{Bernoulli}(p)$. Here, p is the probability that an average summary token is hallucinated, computed on training data. MOSTFREQ5 and MOSTFREQ10 examine the semantic types of events (Table 2). If an event type is frequently hallucinated, all of its tokens are labeled as 1. As seen in the table, we find our H-Alert can not only outperform the baselines, but it obtains a high recall score (71.66%).

M-Alert is our missing event detector. It predicts source medical events that are missed by the summary. Baseline-RAND samples a label for each source event, $e_i \sim \text{Bernoulli}(q)$, where q is the probability an average event is missed, computed on training data. We find that M-Alert produces better precision scores than all baselines. The best performance is achieved by the model variant +TYPE, which injects event types to the BigBird model to help detection of missing events. We note that identifying key medical events remains a challenging task and graph neural networks may help model inter-event relations.

Expert P/R/F. On expert-annotated summaries, we report scores for both detectors. System alerts have been manually verified. The micro-averaged P/R/F scores for H-Alert is 17.24/**58.82**/26.66, and the scores for M-Alert is 30.65/**53.84**/39.06. These results are positive because both detectors are able to attain high recall scores, indicating errors could be effectively flagged and passed on to physicians for further review.

6 Discussion

We discuss our findings from interviewing physicians and underline some of the key areas that are indispensable for further progress on this task.

- ***Medical jargon.*** Owing to time constraints and the literacy of physicians who create the clinical notes, the data we received are of varying quality. It is not uncommon to find jargon or ambiguous information, e.g., "*Patient presents w/ < 24 hours abdominal pain nausea and non-bloody V/D*," here, "V/D" refers to "vomit and diarrhea."
- ***Style difference in clinical notes.*** The notes could be: 1) *procedure-oriented*, i.e., they are narratives describing medical procedures performed on the patient, including treatment, medication, care plans and etc. 2) *disease-oriented*, i.e., each of the patient's diseases is addressed in a separable section, or 3) *organ-oriented*, i.e., each organ is addressed in a separable section.
- ***Improper grounding.*** An after-visit summary states "*We did test you for the coronavirus which was negative*." However, the "*coronavirus test*" was nowhere to be found in the source document. Similar grounding issue was identified in 5 out of 18 summaries during expert revision. Sometimes physicians directly include their knowledge about the patients into after-visit summaries without referring to clinical notes, causing a summarizer fine-tuned on such data to also "hallucinate" content.
- ***High variance in length.*** It would be unwise to truncate clinical notes, despite that most neural models use a fixed maximum length. E.g., a patient who underwent a heart transplant has a high risk of multiple medical comorbidities. It can lead to a large volume of EHR notes. Interestingly, physicians tend to include *more* content in after-visit summaries if they believe patients have *high medical literacy* and are able to understand and act upon complex instructions. This indicates that future systems may produce summaries of varying length per patients' needs.

A System Generated Summary: You were admitted for dizziness. You had a CT scan of your head which showed some thickening in the sinuses of your sinuses. You were seen by the ear nose and throat doctor who recommended that you take an antibiotic called Unasyn while you are in the hospital. You also had an MRI of your brain which did not show any stroke. You are doing better and can go home today.
After Physician's Revision: You were admitted for dizziness. You had a CT scan of your head which showed some thickening in your sinuses and mastoid. This could be suggestive of an infection but your white cells and temperature were normal. You were seen by the ear nose and throat doctor who recommended that you take an antibiotic called Unasyn while you are in the hospital. You also had an MRI of your brain which did not show any stroke. You are doing better and can go home today.

Table 8: A direct comparison of summaries before and after physician revision. A post-study interview with physicians reveals that most revisions are related to missing key medical events (colored orange). They also spend substantial efforts explaining medical jargon to patients and fixing hallucinations (colored red).

7 Conclusion

We tackle the problem of generation of patient after-visit summaries. We compared state-of-the-art summarization models for this task and introduced a novel alerting mechanism to predict two types of errors, including missing medical events and hallucinations in summaries. Extensive experiments using automatic metrics and expert evaluation show the effectiveness of our proposed approach.

8 Ethical Considerations

Data. Data used in this study are obtained from a comprehensive inpatient medical facility. They are electronic dismissal notes created by physicians to record a patient's hospital stay or a series of treatments performed on a patient. These EHR notes are information-dense and full of technical terms. They need to be rewritten and summarized to generate after-visit summaries. The purpose of using patient medical records is to fine-tune abstractive summarization systems and quantitatively evaluate the truthfulness and adequacy of system summaries. These medical records are not for non-academic uses and intents. All medical records are deidentified by the hospital to protect patient privacy.

Summarization Models. Models for abstractive summarization have a tendency to hallucinate information that is not present in the input documents. This is because abstractive models carry inductive biases rooted in the data they are pretrained on. The data encode prior knowledge of natural language, they may also contain a non-negligible amount of toxic and abusive content. Despite our best efforts to alert clinicians of potential errors, some of them could be almost unnoticeable by non-physicians. We thus caution our users to carefully consider the ethical issues specific to abstractive summarization and natural language generation models.

Acknowledgement

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number NIH R01LM012817. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Fei Liu is supported in part by National Science Foundation grant IIS-1909603.

References

Sabita Acharya, Barbara Di Eugenio, Andrew Boyd, Richard Cameron, Karen Dunn Lopez, Pamela Martyn-Nemeth, Carolyn Dickens, and Amer Ardati. 2018. Towards generating personalized hospitalization summaries. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 74–82, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What's in a summary? laying the groundwork for advances in hospital-course summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4794–4811, Online. Association for Computational Linguistics.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview of the first workshop on scholarly document processing (SDP). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 1–6, Online. Association for Computational Linguistics.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Barbara Di Eugenio, Andrew Boyd, Camillo Lugaresi, Abhinaya Balasubramanian, Gail Keenan, Mike Burton, Tamara Goncalves Rezende Macieira, Jianrong Li, Yves Lussier, and Yves Lussier. 2014. Patient-Narr: Towards generating patient-centric summaries of hospital stays. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 6–10, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Alex Federman, Erin Sarzynski, Cindy Brach, Paul Francaviglia, Jessica Jacques, Lina Jandorf, Angela Sanchez Munoz, Michael Wolf, and Joseph Kannry. 2018. Challenges optimizing the after visit summary. *International journal of medical informatics*, 120:14–19.

Jemima A Frimpong, Christopher G Myers, Kathleen M Sutcliffe, and Yemeng Lu-Myers. 2017. When health care providers look at problems from multiple perspectives, patients benefit. *Harv Bus Rev. June*, 23.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Jamie S Hirsch, Jessica S Tanenbaum, Sharon Lipsky Gorman, Connie Liu, Eric Schmitz, Dritan Hashorva, Artem Ervits, David Vawdrey, Marc Sturm, and Noémie Elhadad. 2015. Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274.

Judith Hong, Tien V Nguyen, and Neil S Prose. 2013. Compassionate care: Enhancing physician–patient communication and education in dermatology: Part ii: Patient education. *Journal of the American Academy of Dermatology*, 68(3):364–e1.

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*, pages 3755–3763, Online. Association for Computational Linguistics.

Kundan Krishna, Sopan Khosla, Jeffrey 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 (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Reno Kriz, Marianna Apidianaki, and Chris Callison-Burch. 2020. Simple-qe: Better automatic quality estimation for text simplification. *arXiv preprint arXiv:2012.12382*.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3137–3147, Minneapolis, Minnesota. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Mark Kutner, Elizabeth Greenburg, Ying Jin, and Christine Paulsen. 2006. The health literacy of america's adults: Results from the 2003 national assessment of adult literacy. nces 2006-483. *National Center for Education Statistics*.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. Association for Computational Linguistics.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020. Understanding points of correspondence between sentences for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 191–198, Online. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation.

Ansong Ni, Zhangir Azerbayev, Mutethia Mutuma, Troy Feng, Yusen Zhang, Tao Yu, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. Summertime: Text summarization toolkit for non-experts.

Kevin J O'Leary, Nita Kulkarni, Matthew P Landler, Jiyeon Jeon, Katherine J Hahn, Katherine M Englert, and Mark V Williams. 2010. Hospitalized patients' understanding of their plan of care. In *Mayo Clinic Proceedings*, volume 85, pages 47–52. Elsevier.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Sarita Pathak, Gregory Summerville, Celia P Kaplan, Sarah S Nouri, and Leah S Karliner. 2020. Patient-reported use of the after visit summary in a primary care internal medicine practice. *Journal of Patient Experience*, 7(5):703–707.

Preethi Raghavan, Eric Fosler-Lussier, and Albert Lai. 2012. Temporal classification of medical events. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 29–37, Montréal, Canada. Association for Computational Linguistics.

Ruth Reátegui and Sylvie Ratté. 2018. Comparison of metamap and ctakes for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making*, 18(3):74.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

H. Singh, A. N. Meyer, and E. J. Thomas. 2014. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. *BMJ Qual Saf*, 23(9):727–731.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1899–1905, Online. Association for Computational Linguistics.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 79–85, New Orleans, Louisiana. Association for Computational Linguistics.

Colin P West, Liselotte N Dyrbye, and Tait D Shanafelt. 2018. Physician burnout: contributors, consequences and solutions. *Journal of internal medicine*, 283(6):516–529.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Manzil Zaheer, Guru Guruganesh, Kumar 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. *Advances in Neural Information Processing Systems*, 33.

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

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

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, Brussels, Belgium. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

A Appendix

A.1 Implementation Details

Extractive Summarizers:

- A extractive summary contains 8 sentences, they are generated by LexRank, TextRank, BertSum.
- LexRank/TextRank source code: SummerTime
- We use default settings of BertSum for training. The source code and configs are available here.

Abtractive Summarizers:

- We implement our models based on Huggingface Transformers
- Configurations of our abstractive models: num train epochs: 6; max target length: 256; max source length: 1024; batch size: 2; beam size: 1; topK: 50.
- The models we explored: facebook/bart-large; google/pegasus-large; allenai/led-large-16384

Type I Error Detector (Hallucination):

- We implement Big-Bird based on Huggingface Transformers's BigBird implementation
- Key parameters of the model are - max sequence length: 1536; num train epochs 3; batch size: 4. For other hyper-parameters we use the system's default setting.
- The model is: google/bigbird-roberta-base

Type II Error Detector (Missing Content):

- We implement Big-Bird based on Huggingface Transformers's BigBird implementation
- Key parameters of the model are - max sequence length: 1024; num train epochs 3; batch size: 4. For other hyper-parameters we use the system's default setting.
- The model is: google/bigbird-roberta-base
- MetaMap-2020 version is used for medical event identification.

Summarization Evaluation:

- We use rouge-score to obtain ROUGE scores.
- We use SummerTime to obtain BertScores.
- We use this script for SARI evaluation.
- We use py-readability-metrics to compute Dale Chall readability scores.

All Huggingface models could be downloaded from the Huggingface website. All the models are trained and tested on a NVIDIA-V100 GPU. The average training time for our generation models is around 6-8 hours, the average training time for our error alerting models is around 3-5 hours.

Original	Nuclear stress test which was negative for any damage to the heart.
Synthesized	CT scan of the head which was negative for any damage to the brain.
Original	During hospitalization you underwent endovascular repair of your thoracoabdominal aortic aneurysm with Dr [**NAME**]
Synthesized	During hospitalization you underwent a biopsy of your liver with Dr [**NAME**]
Original	You underwent MRI imaging that showed no obstruction in your GI tract.
Synthesized	You underwent cardiac catheterization that showed no abnormalities.
Original	You were found to have decreased levels of oxygen in your blood.
Synthesized	You were found to have bacteria in your blood.
Original	You were admitted for high calcium levels in your blood.
Synthesized	You were admitted for pain in your left leg.

Table 9: Example synthesized summary sentences that contain hallucinations (underlined).

A.2 Example Outputs

We present example source documents, system and reference summaries, and physician-edited system summaries in Tables 8 and 10.

A.3 Human Evaluation Details

The mean Pearson's r of human evaluation scores is 0.282, suggesting a moderate correlation. While the human score correlation is not strong, it is not surprising as studies have revealed that physicians often have vastly different ways of seeing and treating patients, as differences in profession, specialty, experience, or background lead them to pay attention to particular signals or cues and influence how they approach problems (Frimpong et al., 2017). Specifically, we observe when scoring the readability of generated AVS notes, some physician think some complicated phrases (e.g. CT Scan, mental status) would impact patients' understanding, other physicians would think these phrases would be acceptable to be contained in an after-visit summary.

Example 1
Source Document: [** MISC **]y o female with history and exam consistent with left pyelonephritis. RENAL/ID: Patient presented to ED with fever left CVA tenderness and UA consistent with pyelonephritis. During her hospital stay she was treated with IV fluids zofran toradol tylenol and given 2 dose of ceftriaxone. She was discharged on Bactrim 400 mg BID for a total of 14 days. As she already had 2 doses prior to admission she will continue with Bactrim for 12 more days. She was instructed to follow up with PCP within 1 week. FEN/GI: Patient with some nausea on presentation. Will continue with regular diet and discharged with PO Zofran for nausea as needed. NEURO: Patient with back pain secondary to pyelonephritis. She will continue with Tylenol and Ibuprofen as needed. [...]
After-Visit Summary (Reference): You were admitted because you were having abdominal pain. This was initially thought to be due to a kidney infection but an ultrasound showed that you had a cyst in your left ovary. You went to the operating room for a diagnostic laparoscopy and cystectomy. After the operation you had a couple of fainting episodes and your blood levels were decreasing. Your abdominal exam was also concerning. You were taken back to the operating room and were found to have bleeding in your abdomen; the blood was removed and no active bleeding was noted. You tolerated the second procedure well and were able to eat void and your pain was controlled. You will be discharged home in stable condition with follow up with Dr. [** PERSON **].
After-Visit Summary (System): You came to the hospital because you were having fevers and back pain. You were found to have a kidney infection that spread up to your kidneys. We gave you IV fluids and antibiotics and your symptoms improved. You are being discharged with instructions to continue to take oral antibiotics for 12 more days. Please follow up with your PCP within 1 week.
After-Visit Summary (System w/ Physician Edits): You came to the hospital because you were having fevers and back pain. You were found to have a urine infection that spread up to your kidneys. We gave you IV fluids and antibiotics and your symptoms improved. You are being discharged with instructions to continue to take oral antibiotics for 12 more days and PO Zofran for nausea as needed. Please follow up with your PCP within 1 week.
Example 2
Source Document: Mastoiditis Patient presents with symptoms of vertigo. CT head showed that there is mucosal thickening at the bilateral maxillary sinuses and minimal fluid in the bilateral mastoid air cells. Unclear if this is acute or chronic. Patient describes feeling of fullness in head and decrease in hearing over one month. No fever no leukocytosis. Concern for infection and possible mastoiditis. Dr. [** PERSON **] from *[anonymized]* was consulted by ED recommends Unasyn while inpatient Augmentin as an outpatient for 1 week and will follow up as outpatient. Vertigo History of vertigo and on meclizine as outpatient. No improvement with home dose of meclizine. CT scan findings as detailed above. No intracranial mass bleed. Given IVF in ED. given antiemetics and meclizine and treat for possible mastoiditis /mastoid effusions. Neurology dr. [** PERSON **] consulted. Doubt TIA. Likely peripheral nystagmus and MRI head showed no CVA. Essential (primary) hypertension Controlled. Continue atenolol at 25mg nightly and amlodipine 5mg daily. [...]
After-Visit Summary (Reference): You were admitted with vertigo. The head CT showed no acute abnormalities of the brain. The MRi of brain showed only chronic changes. No acute abnormalities. Your head CT showed possible mastoidit is and thickening of the lining of your maxillary sinuses which could be sinusitis related. ENT was called and recommended iv antibiotics for one day and to transition over to oral antibiotics to complete a total course of one week and you will follow up with [** PERSON **] in one week. You were seen by neurology as well. You are doing much better and can go home today.
After-Visit Summary (System): You were admitted for dizziness. You had a CT scan of your head which showed some thickening in the sinuses of your sinuses. You were seen by the ear nose and throat doctor who recommended that you take an antibiotic called Unasyn while you are in the hospital. You also had an MRI of your brain which did not show any stroke. You are doing better and can go home today.
After-Visit Summary (System w/ Physician Edits): You were admitted for dizziness. You had a CT scan of your head which showed some thickening in your sinuses and mastoid. This could be suggestive of an infection but your white cells and temperature were normal. You were seen by the ear nose and throat doctor who recommended that you take an antibiotic called Unasyn while you are in the hospital. You also had an MRI of your brain which did not show any stroke. You are doing better and can go home today.

Table 10: Example input documents, system and reference summaries, as well as physician-edited system summaries.