What's in a Summary?
Laying the Groundwork for Advances in Hospital-Course Summarization

Griffin Adams[1], **Emily Alsentzer**[2], **Mert Ketenci**[1], **Jason Zucker**[1], **Noémie Elhadad**[1]

[1]Columbia University, New York, NY

[2]Harvard-MIT's Health Science and Technology, Cambridge, MA

{griffin.adams, mert.ketenci, noemie.elhadad}@columbia.edu

emilya@mit.edu, jz2700@cumc.columbia.edu

Abstract

Summarization of clinical narratives is a long-standing research problem. Here, we introduce the task of hospital-course summarization. Given the documentation authored throughout a patient's hospitalization, generate a paragraph that tells the story of the patient admission. We construct an English, text-to-text dataset of 109,000 hospitalizations (2M source notes) and their corresponding summary proxy: the clinician-authored "Brief Hospital Course" paragraph written as part of a discharge note. Exploratory analyses reveal that the BHC paragraphs are highly abstractive with some long extracted fragments; are concise yet comprehensive; differ in style and content organization from the source notes; exhibit minimal lexical cohesion; and represent silver-standard references. Our analysis identifies multiple implications for modeling this complex, multi-document summarization task.

1 Introduction

The electronic health record (EHR) contains critical information for clinicians to assess a patient's medical history (e.g., conditions, laboratory tests, procedures, treatments) and healthcare interactions (e.g., primary care and specialist visits, emergency department visits, and hospitalizations). While medications, labs, and diagnoses are documented through structured data elements and flowsheets, clinical notes contain rich narratives describing the patient's medical condition and interventions. A single hospital visit for a patient with a lengthy hospital stay, or complex illness, can consist of hundreds of notes. At the point of care, clinicians already pressed for time, face a steep challenge of making sense of their patient's documentation and synthesizing it either for their own decision making process or to ensure coordination of care (Hall and Walton, 2004; Ash et al., 2004).

Automatic summarization has been proposed to support clinicians in multiple scenarios, from making sense of a patient's longitudinal record over long periods of time and multiple interactions with the healthcare system, to synthesizing a specific visit's documentation. Here, we focus on *hospital-course summarization*: faithfully and concisely summarizing the EHR documentation for a patient's specific inpatient visit, from admission to discharge. Crucial for continuity of care and patient safety after discharge (Kripalani et al., 2007; Van Walraven et al., 2002), hospital-course summarization also represents an incredibly challenging multi-document summarization task with diverse knowledge requirements. To properly synthesize an admission, one must not only identify relevant problems, but link them to symptoms, procedures, medications, and observations while adhering to temporal, problem-specific constraints.

Our main contributions are as follows: (1) We introduce the task of hospital-course summarization; (2) we collect a dataset of inpatient documentation and corresponding "Brief Hospital Course" paragraphs extracted from discharge notes; and (3) we assess the characteristics of these summary paragraphs as a proxy for target summaries and discuss implications for the design and evaluation of a hospital-course summarization tool.

2 Related Works

Summarization of clinical data and documentation has been explored in a variety of use cases (Pivovarov and Elhadad, 2015). For longitudinal records, graphical representations of structured EHR data elements (i.e., diagnosis codes, laboratory test measurements, and medications) have been proposed (Powsner and Tufte, 1997; Plaisant et al., 1996). Interactive visualizations of clinical problems' salience, whether extracted from notes (Hirsch et al., 2015) or inferred from clinical documentation (Levy-Fix et al., 2020) have shown promise (Pivovarov et al., 2016; Levy-Fix, 2020).

Most work in this area, however, has focused on

4794

clinical documentation of a fine temporal resolution. Traditional text generation techniques have been proposed to synthesize structured data like ICU physiological data streams (Hunter et al., 2008; Goldstein and Shahar, 2016). Liu (2018) use a transformer model to write EHR notes from the prior 24 hours, while Liang et al. (2019) perform disease-specific summarization from individual progress notes. McInerney et al. (2020) develop a distant supervision approach to generate extractive summaries to aid radiologists when interpreting images. Zhang et al. (2018, 2020); MacAvaney et al. (2019); Sotudeh Gharebagh et al. (2020) generate the "Impression" section of the Radiology report from the more detailed "Findings" section. Finally, several recent works aim to generate EHR notes from doctor-patient conversations (Krishna et al., 2020; Joshi et al., 2020; Research, 2020). Recent work on summarizing hospital admissions focuses on extractive methods (Moen et al., 2014, 2016; Liu et al., 2018b; Alsentzer and Kim, 2018).

3 Hospital-Course Summarization Task

Given the clinical documentation available for a patient hospitalization, our task of interest is to generate a text that synthesizes the hospital course in a faithful and concise fashion. For our analysis, we rely on the "Brief Hospital Course" (BHC), a mandatory section of the discharge note, as a proxy reference. The BHC tells the story of the patient's admission: *what* was done to the patient during the hospital admission and *why*, as well as the *follow up* steps needed to occur post discharge, whenever needed. Nevertheless, it is recognized as a challenging and time consuming task for clinicians to write (Dodd, 2007; UC Irvine Residency, 2020).

3.1 Dataset

To carry out our analysis, we construct a large-scale, multi-document summarization dataset, CLINSUM. Materials come from all hospitalizations between 2010 and 2014 at Columbia University Irving Medical Center. Table 1 shows summary statistics for the corpus. There are a wide range of reasons for hospitalizations, from life-threatening situations (e.g., heart attack) to when management of a specific problem cannot be carried out effectively outside of the hospital (e.g., uncontrolled diabetes). This contributes to the high variance in documentation. For reference, Table 7 provides a comparison of basic statistics to widely used summarization

	Variable	Value	STD
Global	# Patients	68,936	
	# Admissions	109,726	N/A
	# Source Notes	2,054,828	
Per Adm.	Length of Stay	5.8 days	9.0
	# Source Notes	18.7	30.1
	# Source Sentences	1,061.2	1,853.6
	# Source Tokens	11,838.7	21,506.5
	# Summary Sentences	17.8	16.9
	# Summary Tokens	261.9	233.8
Per Sent.	# Source Tokens	10.9	12.4
	# Summary Tokens	14.5	11.5
Ratio	Word Compression	42.5	164.6

Table 1: Basic Statistics for CLINSUM. Value is the total for Global, and average for 'Per Admission' and 'Per Sentence'. STD is standard deviation.

datasets. Relatively speaking, CLINSUM is remarkable for having a very high compression ratio despite having long reference summaries. Additionally, it appears highly extractive with respect to fragment density (we qualify this in Section 4.1).

Based on advice from clinicians, we rely on the following subset of note types as source documents: "Admission", "Progress", and "Consult" notes. The dataset does not contain any structured data, documentation from past encounters, or other note types (e.g., nursing notes, social work, radiology reports) (Reichert et al., 2010). Please refer to Appendix A for more details and rationale.

3.2 Tools for Analysis

Entity Extraction & Linking. We use the Med-CAT toolkit (Kraljevic et al., 2020) to extract medical entity mentions and normalize to concepts from the UMLS (Unified Medical Language System) terminology (Bodenreider, 2004). To exclude less relevant entities, we only keep entities from the Disorders, Chemicals & Drugs, and Procedures semantic groups, or the Lab Results semantic type.

Local Coherence. We examine inter-sentential coherence in two ways. **Next-Sentence Prediction (NSP)**. Since we compare across a few datasets representing different domains, we use domain-specific pre-trained BERT models via HuggingFace (Wolf et al., 2019): "bert-base-cased" for CNN/DM and Arxiv, "monologg/biobert_v1.1_pubmed" for Pubmed, and "emilyalsentzer/Bio_ClinicalBERT" for CLINSUM. **Entity-grids.** Entity-grids model local coherence by considering the distribution of discourse entities (Barzilay and Lapata, 2005). An entity grid is a 2-D representation of a text whose entries represent the presence or absence of a

discourse entity in a sentence. For our analyses, we treat UMLS concepts as entities and train a neural model, similar to Tien Nguyen and Joty (2017); Joty et al. (2018), which learns to rank the entity grid of a text more highly than the same entity grid whose rows (sentences) have been randomly shuffled. Please see Appendix B for more details.

Lexical Overlap Metric. We use ROUGE-1 (R1) & ROUGE-2 (R2) F-1 (Lin, 2004) to measure lexical overlap, while ignoring higher order variants based on analysis from other work (Krishna et al., 2021). We denote the average of R1 & R2 scores as R_{12}.

Extractive Summarization Baselines. We rely on a diverse set of sentence extraction methods, whose performance on a held-out portion of CLIN-SUM is reported in Table 2. **Oracle models** have access to the ground-truth reference and represent upper bounds for extraction. Here, we define the sentence selection criteria for each oracle variant, leaving more in-depth discussion to the subsequent analysis. ORACLE TOP-K: Take sentences with highest R_{12} vis-a-vis the reference until a target token count is reached; ORACLE GAIN: Greedily take source sentence with highest relative R_{12} gain conditioned on existing summary[1]. Extract sentences until the change in R_{12} is negative; OR-ACLE SENT-ALIGN: For each sentence in reference, take source sentence with highest R_{12} score; ORACLE RETRIEVAL: For each sentence in reference, take reference sentence from train set with largest BM25 score (Robertson and Walker, 1994); and ORACLE SENT-ALIGN + RETRIEVAL: For each sentence in reference, take sentence with highest R_{12} between ORACLE SENT-ALIGN and ORA-CLE RETRIEVAL. We provide two **unsupervised methods** as well. RANDOM: extracts random sentences until summary reaches target word count (average summary length); LEXRANK: selects the top-k sentences with largest LexRank (Erkan and Radev, 2004) score until target word count is reached. For a supervised baseline, we present CLINNEUSUM: a variant of the Neusum model adapted to the clinical genre (Zhou et al., 2018). CLINNEUSUM is a hierarchical LSTM network trained on ground-truth labels derived from ORA-CLE GAIN, which we detail in Appendix C.

---

[1] This is the Neusum model's objective (Zhou et al., 2018)

4 Dataset Analysis & Implications

To motivate future research in multiple, self-contained directions, we distill task-specific characteristics to a few salient, standalone takeaways. For each takeaway, we provide evidence in the data and/or literature, before proposing implications of findings on model development and evaluation.

4.1 Summaries are mostly abstractive with a few long segments of copy-pasted text

tl;dr. CLINSUM summaries appear extractive according to widely used metrics. Yet, there is large variance within summaries. This directly affects the performance of a supervised extractive model, whose selection capability degrades as summary content transitions from copy-paste to abstractive. In turn, we need models which can handle abrupt transitions between extractive and abstractive text.

Background. Clinicians copy forward information from previous notes to save time and ensure that each note includes sufficient evidence for billing and insurance purposes (Wrenn et al., 2010). Copy-paste is both widely used (66-90% of clinicians according to a recent literature review (Tsou et al., 2017)) and widely applied (a recent study concluded that in a typical note, 18% of the text was manually entered; 46%, copied; and 36% imported[2] (Wang et al., 2017)). Please see Appendix D for more information on the issue of copy-paste.

Analysis - extractiveness. CLINSUM appears very extractive: a high coverage (0.83 avg / 0.13 std) and a very high density (13.1 avg / 38.0 std) (See Grusky et al. (2018) for a description of the statistics). However, we find that 64% of the extractive fragments are unigrams, and 25% are bigrams, which indicate a high level of re-writing. The density measure is large because the remaining 11% of extractive fragments are very long.

Yet, there is a strong positional bias within summaries for long fragments. **Figure 1**, groups fragments according to their relative order within each summary. The longest fragments are usually first. Qualitative analysis confirms that the beginning of the BHC is typically copied from a previous note and conveys the "one-liner" (e.g., *pt is a 50yo male with history of CHF who presents with edema.*)

This abrupt shift in extractiveness should affect content selection. In particular, when look-

---

[2] Imported refers to text typically pulled in from structured data, such as a medication or problem list.

Extractive Baseline	ROUGE-1			ROUGE-2		
	Recall	Precision	F1	Recall	Precision	F1
RANDOM	0.16	0.24	0.17	0.04	0.03	0.03
LEXRANK	0.18	0.21	0.18	0.05	0.05	0.05
CLINNEUSUM	0.36	0.25	0.27	0.14	0.1	0.11
ORACLE TOP-K	0.28	0.52	0.32	0.16	0.32	0.19
ORACLE GAIN	0.43	0.63	0.5	0.26	0.42	0.3
ORACLE SENT-ALIGN (SA)	0.48	0.61	0.52	0.3	0.33	0.31
ORACLE RETRIEVAL	0.51	0.70	0.58	0.25	0.28	0.29
ORACLE SA + RETRIEVAL	**0.6**	**0.76**	**0.66**	**0.4**	**0.49**	**0.43**

Table 2: Performance of different sentence selection strategies on CLINSUM.
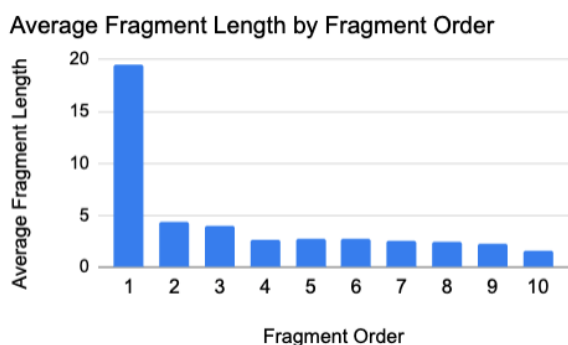
Figure 1: Average extractive fragment lengths according to their relative order within the summary.
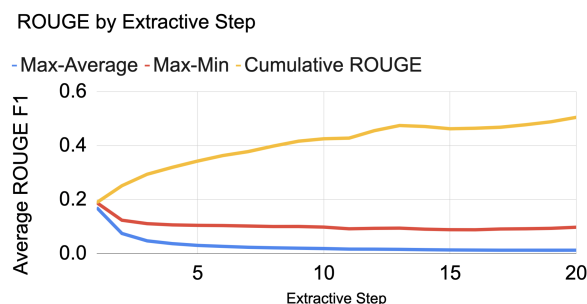

Figure 2: We plot average ROUGE score as summaries are greedily built by adding the sentence with the highest relative ROUGE gain vis-a-vis the current summary, until the gain is no longer positive (ORACLE GAIN). We also include the difference between the highest scoring sentence and the average / minimum to demonstrate a weakening sentence selection signal after the top 1-2.

ing at oracle extractive strategies, we should see clear-cut evidence of (1) 1-2 sentences which are easy to identify as salient (i.e., high lexical overlap with source due to copy-paste), (2) a murkier signal thereafter. To confirm this, we analyze the sentences selected by the ORACLE GAIN method, which builds a summary by iteratively maximizing the R_{12} score of the existing summary vis-a-vis the reference.

In Figure 2, two supporting trends emerge. (1) On average, one sentence accounts for roughly 50%[3] of the overall R_{12} score. (2) Afterwards, the marginal contribution of the next shrinks, as well as the R_{12} gap between the best sentence and the minimum / average, according to the oracle.

There should also be evidence of the copy-paste positional bias impacting content selection. Table 3 reveals that the order in which the ORACLE GAIN summary is built–by maximal lexical overlap with the partially built summary–roughly corresponds to the true ordering of the summary. More simply, the summary transitions from extractive to abstractive.

Unsurprisingly, a model (CLINNEUSUM) trained on ORACLE GAIN extractions gets progressively worse at mimicking it. Specifically, for each extractive step, there exists a ground-truth ranking of candidate sentences by relative R_{12} gain. As the relevance gap between source sentences shrinks (from Figure 2), CLINNEUSUM's predictions deviate further from the oracle rank (Table 4).

Analysis - Redundancy. Even though we prevent all baseline methods from generating duplicate sentences (23% of source sentences have exact match antecedents), there is still a great deal of redundancy in the source notes (i.e., modifications to copy-pasted text). This causes two issues related to content selection. The first is fairly intuitive - that local sentence extraction propagates severe redundancy from the source notes into the summary and, as a result, produces summaries with low lexical coverage. We confirm this by examining the performance between the ORACLE TOP-K and OR-ACLE GAIN, which represent summary-unaware and summary-aware variants of the same selection

---

[3]From Table 2, the average R_{12} score is 0.39 for ORACLE GAIN. To reconcile this number with respect to Figure 2, we note that the average oracle summary is far less than the 20 sentence upper bound shown in the chart.

Extractive Step	Average Rank of Closest Reference Sentence
1	4.7
2	6.0
3	6.3
4	6.7
5	7.3
> 5	10.1

Table 3: ORACLE GAIN greedily builds summaries by repeatedly selecting the sentence which maximizes the R_{12} score of the partially built summary. By linking each extracted sentence to its closest in the reference, we show that this oracle order is very similar to the true ordering of the summary.

Extractive Step	Ground Truth Rank	
	Average	Median
1	28	7
2	69	22
3	74	31
4	79	39
5	76	42
> 5	80	60

Table 4: Rank of selected sentence vis-a-vis oracle rank at each extraction step. A perfectly trained system would have a ground-truth of 1 at each step.

method. While both extract sentences with the highest R_{12} score, ORACLE GAIN outperforms because it incorporates redundancy by considering the relative R_{12} gain from an additional sentence.

The second side effect is perhaps more surprising, and divergent from findings in summarization literature. For most corpora, repetition is indicative of salience. In fact, methods based on lexical centrality, i.e., TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004), still perform very competitively for most datasets. Yet, for CLINSUM, LexRank barely outperforms a random baseline. Poor performance is not only due to redundance, but also a weak link between lexical centrality and salience. The Pearson correlation coefficient between a sentence's LexRank score and its R_{12} overlap with the reference is statistically significant ($p = 0$) yet weak ($r = 0.29$).

Qualitative analysis reveals two principal reasons, both related to copy-paste and/or imported data. The first relates to the propagation of frequently repeated text which may not be useful for summaries: administrative (names, dates), imported structured data, etc. The second relates to sentence segmentation. Even though we use a cus-

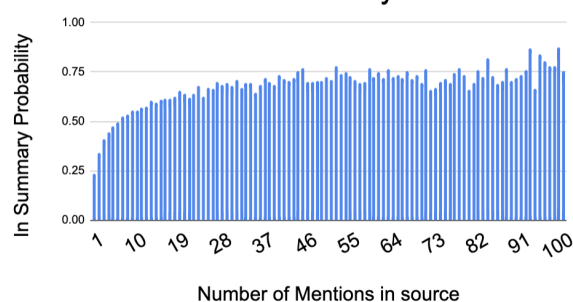## Effect of # Mentions on Entity Salience

Figure 3: Relationship between source entity mentions and probability of inclusion in the summary.

tom sentence splitter, our notes still contain some very long sentences due to imported lists and semi-structured text–a well-documented issue in clinical NLP (Leaman et al., 2015). LexRank summaries have a bias toward these long sentences (26.2 tokens versus source average of 10.9), which have a greater chance of containing lexical centroid(s).

To bypass some of these issues, however, one can examine the link between centrality and salience at the more granular level of entities. **Figure 3** shows a clear-cut positive correlation between source note mention frequency of UMLS concepts and the probability of being included in the summary.

Implications. Regarding within-summary variation in **extractiveness**, we argue for a hybrid approach to balance extraction and abstraction. One of the most widely-used hybrid approaches to generation is the Pointer-Generator (PG) model (See et al., 2017), an abstractive method which allows for copying (i.e., extraction) of source tokens. Another research avenue explicitly decouples the two. These extract-then-abstract approaches come in different flavors: sentence-level re-writing (Chen and Bansal, 2018; Bae et al., 2019), multi-sentence fusion (Lebanoff et al., 2019), and two-step disjoint extractive-abstracive steps (Mendes et al., 2019).

While highly effective in many domains, these approaches do not consider systematic differences in extractiveness within a single summary. To incorporate this variance, one could extend the PG model to copy pre-selected long snippets of text. This would mitigate the problem of copy mechanisms learning to copy very long pieces of text (Gehrmann et al., 2018) - undesirable for the highly abstractive segments of CLINSUM. Span-level extraction is not a new idea (Xu et al., 2020), but, to our knowledge, it has not been studied much in otherwise abstractive settings. For instance, Joshi et al.

([2020](#)) explore patient-doctor conversation summarization and add a penalty to the PG network for over-use of the generator, yet this does not account for intra-summary extractiveness variance.

Regarding **redundancy**, it is clear that, in contrast to some summarization tasks ([Kedzie et al., 2018](#)), summary-aware content selection is essential for hospital course summarization. Given so much noise, massive EHR and cite-specific pre-processing is necessary to better understand the signal between lexical centrality and salience.

4.2 Summaries are concise yet comprehensive

tl;dr. BHC summaries are packed with medical entities, which are well-distributed across the source notes. As such, relations are often not explicit. Collectively, this difficult task calls for a domain-specific approach to assessing faithfulness.

Analysis - concise We find that summaries are extremely dense with medical entities: 20.9% of summary words are medical UMLS entities, compared to 14.1% in the source notes. On average, summaries contain 26 unique entities whereas the source notes contain 265 — an entity compression ratio of 10 (versus token-level compression of 43).

Analysis - comprehensive. Many summarization corpora exhibit systematic biases regarding where summary content can be found within source document(s) ([Dey et al., 2020](#)). On CLINSUM, we examine the distribution of entities along two dimensions: *macro* considers the differences in entity share across notes, and *micro* considers the differences within each note (i.e., lead bias). **(1) Macro Ordering.** When looking at the source notes one by one, how much *additional* relevant information (as measured by entities present in the summary) do you get from each new note? We explore three different orderings: (1) FORWARD orders the notes chronologically, (2) BACKWARD the reverse, and (3) GREEDY ORACLE examines notes in order of decreasing entity entity overlap with the target. Given the large variation in number of notes per admission, we normalize by binning notes into deciles. **Figure 4** shows that it is necessary to read the entire set of notes despite diminishing marginal returns. One might expect the most recent notes to have the most information, considering present as well as copy-forwarded text. Surprisingly, FORWARD and BACKWARD distributions are very similar. GREEDY ORACLE gets at the level

of information concentration. On average, the top 10% of most informative notes cover just over half of the entities found in the summary. We include absolute and percentage counts in Table 5. **(2) Micro Ordering.** We plot a normalized histogram of summary entities by relative position within the source documents. Figure 5 reveals a slight lead bias, followed by an uptick toward the end. Clinical notes are organized by section: often starting with the past medical history and present illness, and typically ending with the plan for future care. All are needed to write a complete BHC.

Ordering	Avg Notes to Read	
	Number	**Percent**
FORWARD	8.5	0.80
BACKWARD	7.8	0.73
GREEDY ORACLE	5.0	0.50

Table 5: Number of documents necessary to cover all relevant UMLS entities—present in the summary—according to three different ordering strategies. FORWARD orders the notes chronologically, BACKWARD the reverse, and GREEDY ORACLE examines notes in order of decreasing entity overlap with the target.
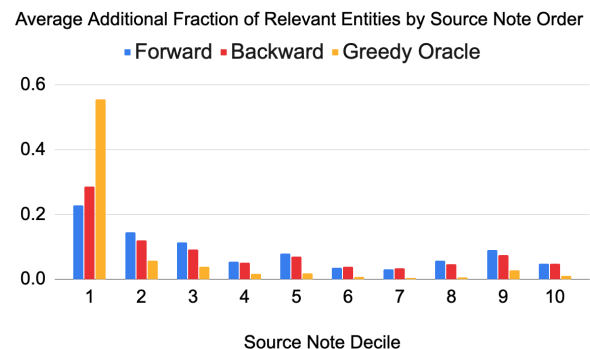
Average Additional Fraction of Relevant Entities by Source Note Order

Figure 4: The average fraction of additional relevant UMLS entities—present in the summary—from reading a patient's visit notes. FORWARD orders the notes chronologically, BACKWARD the reverse, and GREEDY ORACLE in order of decreasing entity overlap.

Implications. The fact that entities are so densely packed in summaries makes models more susceptible to factual errors that misrepresent complex relations. On the CNN/DailyMail dataset, [Goel et al. (2021)](#) reveal performance degradation as a function of the number of entities. This is magnified for clinical text, where failure to identify which treatments were tolerated or discontinued, or to differentiate conditions of the patient or family member, could lead to serious treatment errors.

Recently, the summarization community has explored fact-based evaluation. Yet, many of the
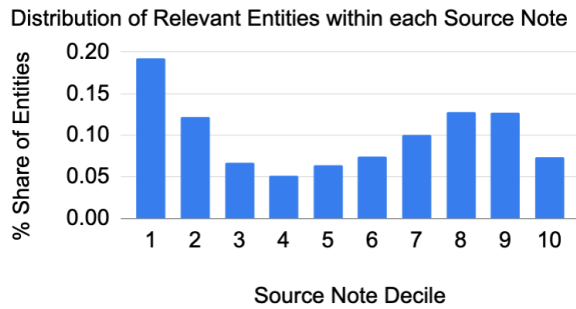
Figure 5: The distribution of relevant entities—present in the summary—within an average source note. Source Note Decile refers to the relative position of each mention within a note. Relevant entities appear throughout an average note, with a slight lead bias.

proposed methods treat global evaluation as the independent sum of very local assessments. In the case of QA-based methods, it is a quiz-like aggregation of individual scores to fairly narrow questions that usually seek to uncover the presence or absence of a single entity or relation. Yet, factoid (Chen et al., 2018), cloze-style (Eyal et al., 2019; Scialom et al., 2019; Deutsch et al., 2020), or mask-conditioned question generation (Durmus et al., 2020) may not be able to directly assess very fine-grained temporal and knowledge-intensive dependencies within a summary. This is a natural byproduct of the fact that many of the factuality assessments were developed for shorter summarization tasks (i.e., headline generation) in the news domain (Cao et al., 2018b; Kryscinski et al., 2019; Maynez et al., 2020). Entailment-based measures to assess faithfulness (Pasunuru and Bansal, 2018; Welleck et al., 2019) can capture complex dependencies yet tend to rely heavily on lexical overlap without deep reasoning (Falke et al., 2019).

Taken together, we argue for the development of fact-based evaluation metrics which encode a deeper knowledge of clinical concepts and their complex semantic and temporal relations[4].

4.3 Summaries have different style and content organization than source notes

tl;dr. Hospital course summarization involves not only massive compression, but a large style and organization transfer. Source notes are written chronologically yet the way clinicians digest the information, and write the discharge summary,

is largely problem-oriented. With simple oracle analysis, we argue that retrieve-edit frameworks are well-suited for hospital course generation.

Analysis - Style. Clinical texts contain many, often obscure, abbreviations (Finley et al., 2016; Adams et al., 2020), misspellings, and sentence fragments (Demner-Fushman et al., 2009). Using a publicly available abbreviation inventory (Moon et al., 2014), we find that abbreviations are more common in the BHC. Furthermore, summary sentences are actually longer on average than source sentences (15.8 versus 12.4 words).

Analysis - Organization. Qualitative analysis confirms that most BHCs are written in a problem-oriented fashion (Weed, 1968), i.e., organized around a patient's disorders. To more robustly analyze content structure, we compare linked UMLS entities at the semantic group level: DRUGS, DISORDERS, and PROCEDURES (McCray et al., 2001). In particular, we compare **global** proportions of semantic groups, **transitions** between entities, as well as **positional** proportions within summaries. **(1) Global.** Procedures are relatively more prevalent in summaries (31% versus 24%), maybe because of the emphasis on events happening during the hospitalization. In both summary and source notes, DISORDERS are the most prevalent (54% and 46%, respectively). Drugs make up 23% and 22% of entity mentions in summary and source notes, respectively. **(2) Transitions.** From both source and summary text, we extract sequences of entities and record adjacent transitions of their semantic groups in a 3×3 matrix. **Figure 7** indicates that summaries have fewer clusters of semantically similar entities (diagonal of the transition matrix). This transition matrix suggests a problem-oriented approach in which disorders are interleaved with associated medications and lab results. **(3) Positional.** Finally, within summaries, we examine the positional relative distribution of semantic groups and connect it to findings from Section 4.1. In **Figure 6**, we first compute the start index of each clinical entity, normalized by the total length, and then group into ten equally sized bins. The early prevalence of disorders and late prevalence of medications is expected, yet the difference is not dramatic. This suggests an HPI-like statement up front, followed by a problem oriented narrative.

If there is a material transfer in **style** and **content**, we would expect that summaries constructed
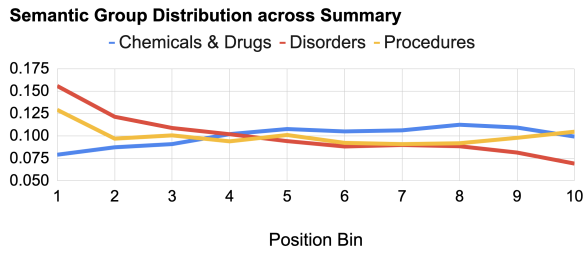
---

[4] Zhang et al. (2020) directly address factuality of clinical text, yet the setting is very different. They explore radiology report accuracy, which is not a temporal multi-document summarization task. Additionally, they rely on a smaller IE system tailored specifically for radiology reports (Irvin et al., 2019).

Semantic Group Distribution across Summary

Figure 6: Position of entities within a summary.

Figure 7: Entity Transition Matrices for source notes and target summaries. Summaries have fewer clusters of semantically similar entities, indicating that entity mentions are woven into a problem-oriented summary.

from other summaries in the dataset would have similar or better lexical coverage than summaries constructed from sentences in the source notes. To assess this, we compare two oracle baselines, SENT-ALIGN and RETRIEVAL. For each sentence in the summary, we find its closest corollary either in the source text (SENT-ALIGN) or in other summaries in the dataset (RETRIEVAL). While the retrieval method is at a distinct disadvantage because it does not contain patient-specific information and retrieval is performed with BM25 scores, we find both methods yield similar results (**Table 2**). An ensemble of SENT-ALIGN and RETRIEVAL performs better than either alone, suggesting that the two types of sources may be complementary. 82% of this oracle's summary sentences are retrievals. Summaries adapt the style and problem-oriented structure of other summaries, but contain patient-specific information from the source notes.

Implications. Hospital-course summaries weave together disorders, medications, and procedures in a problem-oriented fashion. It is clear that substantial re-writing and re-organization of source content is needed. One suitable approach is to use the retrieve-rerank-rewrite (R^3) framework proposed by Cao et al. (2018a). To support this notion, more recent work demonstrates that retrieval augmented generation is effective for knowledge-intensive tasks (Lewis et al., 2020b), enhances sys-
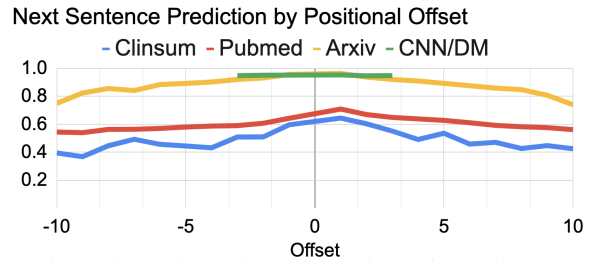

Figure 8: NSP logit by relative position of the next sentence across summaries for several datasets. An offset of 1 corresponds to the true next sentence.

tem interpretability (Guu et al., 2020; Krishna et al., 2020), and can improve LM pre-training (Lewis et al., 2020a)[5]. Also, efforts to bridge the gap between template-based and abstractive generation have been successful in the medical domain for image report generation (Li et al., 2018).

In this light, BHC generation could be truly problem-oriented. The first step would involve selecting salient problems (i.e., disorders) from the source text–a well-defined problem with proven feasibility (Van Vleck and Elhadad, 2010). The second step would involve separately using each problem to retrieve problem-specific sentences from other summaries. These sentences would provide clues to the problem's relevant medications, procedures, and labs. In turn, conceptual overlap could be used to re-rank and select key, problem-specific source sentences. The extracted sentences would provide the patient-specific facts necessary to rewrite the problem-oriented retrieved sentences.

4.4 Summaries exhibit low lexical cohesion

tl;dr. Lexical cohesion is sub-optimal for evaluating hospital-course discourse because clinical summaries naturally exhibit frequent, abrupt topic shifts. Also, low correlation exists between lexical overlap and local coherence metrics.

Analysis. Entity-based coherence research posits that "texts about the same discourse entity are perceived to be more coherent than texts fraught with abrupt switches from one topic to the next" (Barzilay and Lapata, 2005). Yet, for CLINSUM summaries, coherence and abrupt topic shifts are not mutually exclusive. An analysis of the entity grids of summaries, presumably coherent, are sparse, with few lexical chains. In fact, over 66% of the

---

[5]The related idea of template-based generation has gained traction within the probabilistic community (Wiseman et al., 2018; Guu et al., 2018; Wu et al., 2019; He et al., 2020).

entities in the BHC appear only once. Of those with multiple mentions, the percentage which appear in adjacent sentences is only 9.6%. As in Prabhumoye et al. (2020), we also compare coherence with next-sentence prediction (NSP). **Figure 8** plots the NSP logit by positional offset, where an offset of 1 corresponds to the next sentence, and -1 to the previous. NSP relies on word overlap and topic continuity (Bommasani and Cardie, 2020), so it makes sense it is lowest for CLINSUM.

To confirm the hypothesis that ROUGE does not adequately capture content structure, we use the *pairwise ranking* approach to train and evaluate an entity-grid based neural coherence model (Barzilay and Lapata, 2005; Tien Nguyen and Joty, 2017). **Table 6** shows ROUGE and coherence metrics side-by-side for ORACLE GAIN, which naively orders sentences according to document timestamp, then within-document position, and ORACLE SENT-ALIGN, which maintains the structure of the original summary. The poor coherence of ORACLE GAIN is obscured by comparable ROUGE scores.

Summary	Acc.	R1	R2
Actual Summary	0.86	N/A	N/A
ORACLE SENT-ALIGN	0.75	0.52	0.30
ORACLE GAIN	0.54	0.48	0.30

Table 6: Comparison of coherence and ROUGE. Acc. refers to pair-wise ranking accuracy from scoring summaries against random permutations of themselves.

Implications. Content organization is critical and should be explicitly evaluated. A well-established framework for assessing organization and readability is coherence. A large strand of work on modeling coherent discourse has focused on topical clusters of entities (Azzam et al., 1999; Barzilay and Elhadad, 2002; Barzilay and Lee, 2004; Okazaki et al., 2004). Yet, as shown above, CLINSUM summaries exhibit abrupt topic shifts and contain very few repeated entities. The presence and distribution of lexical (Morris and Hirst, 1991; Barzilay and Elhadad, 1997) or co-referential (Azzam et al., 1999) chains, then, might not be an appropriate proxy for clinical summary coherence. Rather, we motivate the development of problem-oriented models of coherence, which are associative in nature, and reflect a deeper knowledge about the relationship between disorders, medications, and procedures. The impetus for task-tailored evaluation metrics is supported by recent meta analyses (Fabbri et al., 2020; Bhandari et al., 2020).

4.5 BHC summaries are silver-standard

tl;dr. Discharge summaries and their associated BHC sections are frequently missing critical information or contain excessive or erroneous content. Modeling efforts should address sample quality.

Analysis. Kripalani et al. (2007) find that discharge summaries often lack important information including diagnostic test results (33-63% missing) treatment or hospital course (7-22%), discharge medications (2-40%), test results pending at discharge (65%), patient/family counseling (90-92%), and follow-up plans (2-43%). The quality of the reporting decreases as the length of the discharge summary increases, likely due to copy-pasted information (van Walraven and Rokosh, 1999).

These quality issues occur for a number of reasons: (1) limited EHR search functionality makes it difficult for clinicians to navigate through abundant patient data (Christensen and Grimsmo, 2008); (2) multiple clinicians contribute to incrementally documenting care throughout the patient's stay; (3) despite existing guidance for residents, clinicians receive little to no formal instruction in summarizing patient information (Ming et al., 2019); and (4) clinicians have little time for documenting care.

Implications. Noisy references can harm model performance, yet there is a rich body of literature to show that simple heuristics can identify good references (Bommasani and Cardie, 2020) and/or filter noisy training samples (Rush et al., 2015b; Akama et al., 2020; Matsumaru et al., 2020). Similar strategies may be necessary for hospital-course generation with silver-standard data. Another direction is scalable reference-free evaluations (ShafieiBavani et al., 2018; Hardy et al., 2019; Sellam et al., 2020; Gao et al., 2020; Vasilyev et al., 2020).

5 Conclusion

Based on a comprehensive analysis of clinical notes, we identify a set of implications for hospital-course summarization on future research. For modeling, we motivate (**1**) the need for dynamic hybrid extraction-abstraction strategies (4.1); (**2**) retrieval-augmented generation (4.3); and (**3**) the development of heuristics to assess reference quality (4.5). For evaluation, we argue for (**1**) methods to assess factuality and discourse which are associative in nature, i.e., incorporate the complex inter-dependence of problems, medications, and labs (4.2, 4.4); and (**2**) scalable reference-free metrics (4.5).

6 Ethical Considerations

Dataset creation. Our CLINSUM dataset contains protected health information about patients. We have received IRB approval through our institution to access this data in a HIPAA-certified, secure environment. To protect patient privacy, we cannot release our dataset, but instead describe generalizable insights that we believe can benefit the general summarization community as well as other groups working with EHR data.

Intended Use & Failure Modes. The ultimate goal of this work is to produce a summarizer that can generate a summary of a hospital course, and thus support clinicians in this cognitively difficult and time-consuming task. While this work is a preface to designing such a tool, and significant advances will be needed to achieve the robustness required for deployment in a clinical environment, it is important to consider the ramifications of this technology at this stage of development. We can learn from existing clinical summarization deployed (Pivovarov et al., 2016) and other data-driven clinical decision support tools (Chen et al., 2020). As with many NLP datasets, CLINSUM likely contains biases, which may be perpetuated by its use. There are a number of experiments we plan to carry out to identify documentation biases and their impact on summarization according to a number of dimensions such as demographics (e.g., racial and gender), social determinents of health (e.g., homeless individuals), and clinical biases (e.g., patients with rare diseases). Furthermore, deployment of an automatic summarizer may lead to automation bias (Goddard et al., 2012), in which clinicians over rely on the automated system, despite controls measures or verification steps that might be built into a deployed system. Finally, medical practices and EHRs systems constantly change, and this distribution drift can cause models to fail if they are not updated. As the NLP community continues to develop NLP applications in safety-critical domains, we must carefully study how can can build robustness, fairness, and trust into these systems.

Acknowledgements

References

Griffin Adams, Mert Ketenci, Shreyas Bhave, Adler Perotte, and Noémie Elhadad. 2020. Zero-shot clinical acronym expansion via latent meaning cells. In *Machine Learning for Health*, pages 12–40. PMLR.

Reina Akama, Sho Yokoi, Jun Suzuki, and Kentaro Inui. 2020. Filtering noisy dialogue corpora by connectivity and content relatedness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 941–958, Online. Association for Computational Linguistics.

Emily Alsentzer and Anne Kim. 2018. Extractive summarization of ehr discharge notes. *arXiv preprint arXiv:1810.12085*.

Joan S Ash, Marc Berg, and Enrico Coiera. 2004. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *Journal of the American Medical Informatics Association*, 11(2):104–112.

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1999. Using coreference chains for text summarization. In *Coreference and Its Applications*.

Sanghwan Bae, Taeuk Kim, Jihoon Kim, and Sanggoo Lee. 2019. Summary level training of sentence rewriting for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 10–20, Hong Kong, China. Association for Computational Linguistics.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Intelligent Scalable Text Summarization*.

Regina Barzilay and Noemie Elhadad. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 113–120, Boston, Massachusetts, USA. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 9347–9359, Online. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018a. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018b. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2020. Ethical machine learning in health care. *arXiv e-prints*, pages arXiv–2009.

Ping Chen, Fei Wu, Tong Wang, and Wei Ding. 2018. A semantic qa-based approach for text summarization evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4800–4807. AAAI Press.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Tom Christensen and Anders Grimsmo. 2008. Instant availability of patient records, but diminished availability of patient information: a multi-method study of gp's use of electronic patient records. *BMC medical informatics and decision making*, 8(1):1–8.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. 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 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2013. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC bioinformatics*, 14(1):10.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2020. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *arXiv preprint arXiv:2010.00490*.

Alvin Dey, Tanya Chowdhury, Yash Kumar, and Tanmoy Chakraborty. 2020. Corpora evaluation and system bias detection in multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2830–2840, Online. Association for Computational Linguistics.

Kimberley Dodd. 2007. Transitions of care – how to write a "good" discharge summary.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. 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.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. 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 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

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.

Gregory P Finley, Serguei VS Pakhomov, Reed McEwan, and Genevieve B Melton. 2016. Towards comprehensive clinical abbreviation disambiguation using machine-labeled training data. In *AMIA Annual Symposium Proceedings*, volume 2016, page 560. American Medical Informatics Association.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

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.

Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127.

Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the nlp evaluation landscape. *arXiv preprint arXiv:2101.04840*.

Ayelet Goldstein and Yuval Shahar. 2016. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. *Journal of biomedical informatics*, 61:159–175.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Amanda Hall and Graham Walton. 2004. Information overload within the health care system: a literature review. *Health Information & Libraries Journal*, 21(2):102–108.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

Junxian He, Taylor Berg-Kirkpatrick, and Graham Neubig. 2020. Learning sparse prototypes for text generation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

Robert E Hirschtick. 2006. Copy-and-paste. *Jama*, 295(20):2335–2336.

Jim Hunter, Yvonne Freer, Albert Gatt, Robert Logie, Neil McIntosh, Marian Van Der Meulen, François Portet, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. 2008. Summarising complex icu data in natural language. In *Amia annual symposium proceedings*, volume 2008, page 323. American Medical Informatics Association.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 590–597. AAAI Press.

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.

Shafiq Joty, Muhammad Tasnim Mohiuddin, and Dat Tien Nguyen. 2018. Coherence modeling of asynchronous conversations: A neural entity grid approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 558–568, Melbourne, Australia. Association for Computational Linguistics.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, et al. 2020. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *arXiv preprint arXiv:2010.01165*.

Sunil Kripalani, Frank LeFevre, Christopher O Phillips, Mark V Williams, Preetha Basaviah, and David W Baker. 2007. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. *Jama*, 297(8):831–841.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

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

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.

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Gal Levy-Fix. 2020. *Patient Record Summarization Through Joint Phenotype Learning and Interactive Visualization*. Ph.D. thesis, Columbia University.

Gal Levy-Fix, Jason Zucker, Konstantin Stojanovic, and Noémie Elhadad. 2020. Towards patient record summarization through joint phenotype learning in HIV patients. *arXiv preprint arXiv:2003.11474*.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1537–1547.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using EHR data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Minneapolis, Minnesota, USA. 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.

Peter J Liu. 2018. Learning to write notes in electronic health records. *arXiv preprint arXiv:1808.02622*.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018a. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Xiangan Liu, Keyang Xu, Pengtao Xie, and Eric Xing. 2018b. Unsupervised pseudo-labeling for extractive summarization on electronic health records. *arXiv preprint arXiv:1811.08040*.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish A. Talati, and Ross W. Filice. 2019. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1013–1016. ACM.

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.

Alexa T McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(0 1):216.

Denis Jered McInerney, Borna Dabiri, Anne-Sophie Touret, Geoffrey Young, Jan-Willem van de Meent, and Byron C Wallace. 2020. Query-focused ehr summarization to aid imaging diagnosis. *arXiv preprint arXiv:2004.04645*.

Afonso Mendes, Shashi Narayan, Sebastião Miranda, Zita Marinho, André F. T. Martins, and Shay B. Cohen. 2019. Jointly extracting and compressing documents with summary state representations. 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 3955–3966, Minneapolis, Minnesota. 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.

David Ming, Kahli Zietlow, Yao Song, Hui-Jie Lee, and Alison Clay. 2019. Discharge summary training curriculum: a novel approach to training medical students how to write effective discharge summaries. *The clinical teacher*, 16(5):507–512.

Hans Moen, Juho Heimonen, Laura-Maria Murtola, Antti Airola, Tapio Pahikkala, Virpi Terävä, Riitta Danielsson-Ojala, Tapio Salakoski, and Sanna Salanterä. 2014. On evaluation of automatically generated clinical discharge summaries. In *PAHI*, pages 101–114.

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

Sungrim Moon, Serguei Pakhomov, Nathan Liu, James O Ryan, and Genevieve B Melton. 2014. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, 21(2):299–307.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3075–3081. AAAI Press.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada. 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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Naoaki Okazaki, Yutaka Matsuo, and Mitsuru Ishizuka. 2004. Improving chronological sentence ordering by precedence relation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 750–756, Geneva, Switzerland. COLING.

Heather C O'Donnell, Rainu Kaushal, Yolanda Barrón, Mark A Callahan, Ronald D Adelman, and Eugenia L Siegler. 2009. Physicians' attitudes towards copy and pasting in electronic note writing. *Journal of general internal medicine*, 24(1):63–68.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. 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 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

Rimma Pivovarov, Yael Judith Coppleson, Sharon Lipsky Gorman, David K Vawdrey, and Noémie Elhadad. 2016. Can patient record summarization support quality metric abstraction? In *AMIA Annual Symposium Proceedings*, volume 2016, page 1020. American Medical Informatics Association.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Catherine Plaisant, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. 1996. Lifelines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–227.

Seth M Powsner and Edward R Tufte. 1997. Summarizing clinical psychiatric data. *Psychiatric Services*, 48(11):1458–1460.

Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2020. Topological sort for sentence ordering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2783–2792, Online. Association for Computational Linguistics.

Daniel Reichert, David Kaufman, Benjamin Bloxham, Herbert Chase, and Noémie Elhadad. 2010. Cognitive analysis of the summarization of longitudinal patient records. In *AMIA Annual Symposium Proceedings*, volume 2010, page 667. American Medical Informatics Association.

Microsoft Research. 2020. Project empowermd: Medical conversations to medical intelligence.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*, pages 232–241. Springer.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015a. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Alexander M Rush, SEAS Harvard, Sumit Chopra, and Jason Weston. 2015b. A neural attention model for sentence summarization. In *ACLWeb. Proceedings of the 2015 conference on empirical methods in natural language processing*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. 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 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

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

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018. Summarization evaluation in the absence of human model summaries using the compositionality of word embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 905–914, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Eugenia L Siegler and Ronald Adelman. 2009. Copy and paste: a remediable hazard of electronic health records. *The American journal of medicine*, 122(6):495–496.

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.

Dat Tien Nguyen and Shafiq Joty. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, Vancouver, Canada. Association for Computational Linguistics.

Amy Y Tsou, Christoph U Lehmann, Jeremy Michel, Ronni Solomon, Lorraine Possanza, and Tejal Gandhi. 2017. Safe practices for copy and paste in the ehr: systematic review, recommendations, and

novel model for health it collaboration. *Applied clinical informatics*, 8(1):12.

UC Irvine Residency. 2020. Resident guide - note writing inpatient medicine wards.

Tielman T Van Vleck and Noémie Elhadad. 2010. Corpus-based problem selection for ehr note summarization. In *AMIA Annual Symposium Proceedings*, volume 2010, page 817. American Medical Informatics Association.

Carl van Walraven and Ella Rokosh. 1999. What is necessary for high-quality discharge summaries? *American Journal of Medical Quality*, 14(4):160–169.

Carl Van Walraven, Ratika Seth, Peter C Austin, and Andreas Laupacis. 2002. Effect of discharge summary availability during post-discharge visits on hospital readmission. *Journal of general internal medicine*, 17(3):186–192.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

Michael D Wang, Raman Khanna, and Nader Najafi. 2017. Characterizing the source of text in electronic health record progress notes. *JAMA internal medicine*, 177(8):1212–1213.

Lawrence L Weed. 1968. Medical records that guide and teach (concluded). *Yearbook of Medical Informatics*, 212:1.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3174–3187. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Jesse O Wrenn, Daniel M Stein, Suzanne Bakken, and Peter D Stetson. 2010. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17(1):49–53.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7281–7288. AAAI Press.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Thomas R Yackel and Peter J Embi. 2006. Copy-and-paste-and-paste. *JAMA*, 296(19):2315–2316.

Rui Zhang, Serguei Pakhomov, Bridget T McInnes, and Genevieve B Melton. 2011. Evaluating measures of redundancy in clinical texts. In *AMIA annual symposium proceedings*, volume 2011, page 1612. American Medical Informatics Association.

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.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

A Additional Dataset Description

Based on advice from clinicians, we rely on the following subset of notes as source documents: "Admission notes", which convey the past medical history of a patient, ongoing medications, and a detailed description of chief complaint; "Progress notes", which convey a daily report about patient status and care as well as to-do lists for next day;

	Dataset	# Docs	Comp. Ratio	Extractiveness		Summary		Source
				Coverage	Density	# words	# sents	# words
SDS	Gigaword (Rush et al., 2015a)	4mn	3.8	0.58	1.1	8.3	1	31.4
	CNN/DM (Nallapati et al., 2016)	312k	13.0	0.80	3.0	55.6	3.8	789.9
	Newsroom (Grusky et al., 2018)	1.2mn	43.0	0.82	9.6	30.4	1.4	750.9
	XSum (Narayan et al., 2018)	226k	18.8	0.57	0.89	23.3	1.0	431.1
	Arxiv (Cohan et al., 2018)	215k	39.8	0.92	3.7	292.8	9.6	6,913.8
	PubMed (Cohan et al., 2018)	133k	16.2	0.90	5.9	214.4	6.9	3,224.4
	BigPatent (Sharma et al., 2019)	1.3mn	36.4	0.86	2.4	116.5	3.5	3,572.8
MDS	WikiSum (Liu et al., 2018a)	2.3mn	264.0	N/A	N/A	139.4	N/A	36,802.5
	Multi-News (Fabbri et al., 2019)	56k	8.0	0.68	3.0	263.7	10	2,103.5
	SOAP (Krishna et al., 2020)	7k	4.7	N/A	N/A	320	N/A	1,500
	CLINSUM (ours)	110k	45.2	0.83	13.1	261.9	17.7	11,838.7

Table 7: Basic statistics for single-document (SDS) and multi-document (MDS) summarization datasets. For multi-document summarization (MDS), # Source words are aggregated across documents. Compression ratio is the average ratio of source words to summary words. Extractiveness metrics (coverage and density) come from Grusky et al. (2018) and, for consistency, are calculated using the official code across the validation set for each dataset. Spacy tokenization is performed before extracting fragments. Other corpus statistics are pulled from either the corresponding paper or Table 1 in Sharma et al. (2019). Entries are filled with N/A because the dataset is private (Krishna et al., 2020), or too expensive to generate (Liu et al., 2018a). The Gigaword SDS dataset comes from the annotated Gigaword dataset (Graff et al., 2003; Napoles et al., 2012).

and "Consult notes", which document specialist consultations. The dataset does not contain any structured data, documentation from past encounters, or other note types (e.g., nursing notes, social work, radiology reports) (Reichert et al., 2010).[6] Additionally, we remove all visits without at least one source note and at least one Brief Hospital Course target section and exclude notes with less than 25 characters. For computational and modeling feasibility, we bound the minimum and maximum lengths for the source and target texts. We exclude visits where the source notes are collectively over $20,000$ tokens ($< 10\%$ of visits) or are shorter than the Brief Hospital Course. Finally, we exclude visits where the Brief Hospital Course section is less than 25 characters and greater than 500 tokens to remove any incorrectly parsed BHC sections.

B Local Coherence Model Details

The underlying premise of the entity-grid model is that "the distribution of entities in locally coherent texts exhibits certain regularities" (Barzilay and Lapata, 2005). The paper defines entities as coreferent noun phrases, while we use UMLS entities. Additionally, Barzilay and Lapata (2005) add syntactic role information to the grid entries, whereas, without reliable parses, we denote a binary indicator of entity presence. As is common practice, we learn

to rank the entity grid of a text more highly than the same entity grid whose rows (sentences) have been randomly shuffled. Inspired by Joty et al. (2018), we first project the entity grid entries onto a shared embedding space whose vocabulary consists of all the UMLS CUIs and a special <empty> token. As in Tien Nguyen and Joty (2017), we then learn features of original and permuted embedded grids by separately applying 1-D convolutions. Finally, scalars produced by the siamese convolutional networks are used for pairwise ranking.

C CLINNEUSUM Details

As in Nallapati et al. (2017) and Zhou et al. (2018), we extract ground truth extraction labels by greedily selecting sentences which maximize the relative ROUGE gain (R_{12}) from adding an additional sentence to an existing summary. We use basic heuristics to scale the model efficiently to a dataset with such a large set of candidate sentences. To avoid inclusion of sentences with spurious, small relevance based on ROUGE, we filter out extractive steps with a weak learning signal - extractive steps for which either the ROUGE improvement of the highest scoring sentence is less than 1%, or the differential between the least and most relevant sentences is less than 2%. Furthermore, based on manual evaluation, we take steps to reduce the size of the candidate sentence set provided to the model sees during training. First, we de-duplicate sentences and remove sentences with no alphabetical letters or a token count less than 3. Then, we randomly re-

---
[6]We note that most structured data fields are now automatically imported into input notes, and, similarly, findings of reports are available in notes.

move source sentences without any lexical overlap with the summary with a probability determined by source length. This produces a train-test bias, but it is minor because most of the removed sentences are consistently irrelevant (i.e., dates, numerical lists, signature lines, etc.). During training, we randomly sample a single extractive step whose objective is to maximize the KL-Divergence between the model-generated score distribution over sentences and a temperature-smoothed softmax over the relative ROUGE gain.

Similarly to the Neusum model, we employ a simple LSTM-based, hierarchical architecture. We project source and target words onto a shared embedding space. Then, separately, we pass word embeddings to a bi-LSTM sentence encoder. We use the concatenated hidden states from the forward and backward pass as input to another bi-LSTM document encoder. We ignore document boundaries in this setup. We treat the concatenation of the sentence-level hidden state from the sentence encoder and the corresponding hidden state from the document encoder as the final sentence-level representation. Then for each candidate source sentence, we attend to each sentence in the existing summary to compute a summary-aware sentence-representation. Finally, we concatenate both representations and pass through three fully connected layers with Tanh activation. The output is a single scalar score for which we compute the softmax over all candidate sentences. We compare to this distribution to the empirical relative ROUGE distribution[7]. For inference, we greedily extract sentences until the target of 13 sentences (validation average) is reached.

D   A Note on Copy-Paste in Clinical Text

Researchers have explored unintended side effects of copy-paste along many different dimensions: information bloat, reporting errors and incoherence from outdated or inconsistent information (Hirschtick, 2006; Yackel and Embi, 2006; Siegler and Adelman, 2009; O'Donnell et al., 2009; Tsou et al., 2017), and quantifying redundancy (Wrenn et al., 2010; Zhang et al., 2011; Cohen et al., 2013). Quantifying redundancy is non-trivial because copy-paste occurs at different granularities and, quite often, the pasted text is modified. We

do not seek to replicate these studies on CLINSUM. Rather, we examine the impact on summary **extractiveness** and **redundancy**.

---

[7]As in the Neusum model, we first min-max normalize the raw ROUGE gains, and then apply a temperature scalar of 5 before computing the softmax.