

Towards Clinical Encounter Summarization: Learning to Compose Discharge Summaries from Prior Notes

Han-Chin Shing[†], Chaitanya Shivade[‡], Nima Pourdamghani[‡], Feng Nan[‡],
Philip Resnik[†], Douglas Oard[†], Parminder Bhatia[‡]

[†]University of Maryland, [‡]Amazon Web Service AI
{shing, resnik, oard}@umd.edu
{shivadc, nimpourd, nanfen, parmib}@amazon.com

Abstract

The records of a clinical encounter can be extensive and complex, thus placing a premium on tools that can extract and summarize relevant information. This paper introduces the task of generating discharge summaries for a clinical encounter. Summaries in this setting need to be faithful, traceable, and scale to multiple long documents, motivating the use of extract-then-abstract summarization cascades. We introduce two new measures, faithfulness and hallucination rate for evaluation in this task, which complement existing measures for fluency and informativeness. Results across seven medical sections and five models show that a summarization architecture that supports traceability yields promising results, and that a sentence-rewriting approach performs consistently on the measure used for faithfulness (faithfulness-adjusted F_3) over a diverse range of generated sections.

1 Introduction

Clinical notes in the electronic health record (EHR) are used to document the patient’s progress and interaction with clinical professionals. These notes contain rich and diverse information, including but not limited to admission notes, nursing notes, radiology notes, and physician notes. The information the clinicians need, however, is often buried in the sheer amount of text, as the number of clinical notes in an encounter can be in the hundreds. Finding the information can be time-consuming; time that is already in short supply for the clinicians to attend to the patients (Weiner and Biondich, 2006; Sinsky et al., 2016), and can even contribute to the worsening physician burnout crisis (Tawfik et al., 2018; West et al., 2018).

Summarization has the potential to help clinicians make sense of these clinical notes. In this paper, we aim to make progress toward summarizing one of the most common information sources

clinicians interact with — the patient’s clinical encounter. A clinical encounter (Figure 1) documents an interaction between a patient and a healthcare provider (e.g., a visit to the hospital), including structured and unstructured data. Our work focuses on the unstructured clinical notes.

A natural target for summarization is the *discharge summary*: a specialized clinical note meant to be a summary of the clinical encounter, typically written at the time of patient discharge. Each section (e.g., past medical history, brief hospital course, medications on admission) in the discharge summary represents a different aspect of the encounter. By building a system to extract and compose these medical sections from prior clinical notes in the same encounter, we can summarize the information in a format clinicians are already trained to read and understand.

There are significant challenges ahead, however. In this work, we identify three main challenges of summarizing a clinical encounter: (1) an *evidence-based fallback* that allows traceable inspection, (2) the *faithfulness* of the summary, and (3) the *long text* in a clinical encounter. We believe that all three challenges need to be properly addressed before a discussion about deployment can happen. Thus, this work focuses on measuring and understanding how existing state-of-the-art summarization systems perform on these challenges. Additionally, we propose an extractive-abstractive summarization pipeline that directly addresses the *evidence-based fallback* challenge and the *long text* challenge. For the third challenge, *faithfulness*, we introduce an evaluation measure that uses a medical NER system, inspired by recent work on faithfulness in summarization (Maynez et al., 2020; Zhang et al., 2020).

Contributions

- We identify three challenges for summarizing clinical encounters: (1) faithfulness, (2) evidence, and (3) long text.

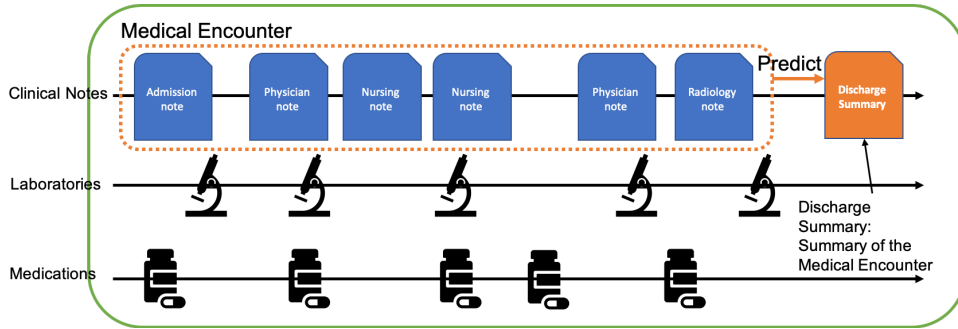


Figure 1: A medical encounter is an interaction between a patient and a healthcare provider.

- We introduce the task of discharge summary composition from prior clinical notes.
- We evaluate our proposed extractive-abstractive pipeline for multi-document summarization with medical NER-based scores and ROUGE across seven discharge summary sections.
- We create a collection derived from a public database (MIMIC III); a potential benchmark for clinical multi-document summarization.

2 Evidence, Faithfulness, and Long Text

In this section, we identify the three main challenges in discharge summary composition.

Evidence. A summary should be displayed with means for the clinician to inspect and understand where the information come from. In this respect, extractive summarization has a clear advantage over abstractive summarization, as the source of extracted content can be easily traced and displayed in context. However, abstractive summarization does benefit from a more fluent generation and the potential to function as a writing aid to alleviate the clinicians’ documentation burden. The challenge lies in how to design the system such that *evidence* can be traced.

Faithfulness. Like any model supporting clinical decision making, measuring and understanding the faithfulness of the model output is important. As abstractive summarization systems are evaluated by their ability to generate fluent output, faithfulness can be a challenge to these models. Addressing this problem is an active area of research (Maynez et al., 2020; Zhang et al., 2020).

Long text. Summarizing an encounter (a sequence of documents), the quantity of text available can easily exceed the memory limit of the model. This memory limitation is especially challenging for modern transformer-based architectures that typically require large GPU-memory. Tokens that

do not fit in memory can contain relevant clinical information for summarization. Attempting to train an abstractive model to generate a summary without the source information available can encourage the model to hallucinate at test time; a dangerous outcome in the context of clinical summarization.

3 Extract and then Abstract

These challenges are common in summarization. In particular, one of the main challenges in multi-document abstractive summarization is to summarize a large number of documents. While significant progress has been made to scale the abstractive models (Beltagy et al., 2020; Zaheer et al., 2020), recent work still involves using an extractive model (e.g., tf-idf based cosine similarity (Liu et al., 2018), logistic regression (Liu and Lapata, 2019a)) to limit the number of paragraphs before abstraction.

Here we proposed a similar extractive-abstractive pipeline. However, what is different in a clinical context is that we wish to place more weight on the extractor than rely on the abstractor to summarize a large amount of text. This decision is motivated by the fact that extractive models are inherently better at being faithful to the source, as they do not introduce novel information. This characteristic makes them ideal candidates for clinical summarization.

Our proposed extractor-abstractor pipeline involves two stages (Figure 2). The first stage functions as a recall-oriented extractive summarization system to extract relevant sentences from prior documents. The extracted sentences are then passed through post-processing steps that remove duplicated sentences and arrange them to form an extractive summary. The second stage is an abstractive summarization system that aims to take the extractive summary from the previous step and smooths out irrelevant or duplicated information. We describe the details of implementations and how to scale this pipeline to very long text in Section 7.

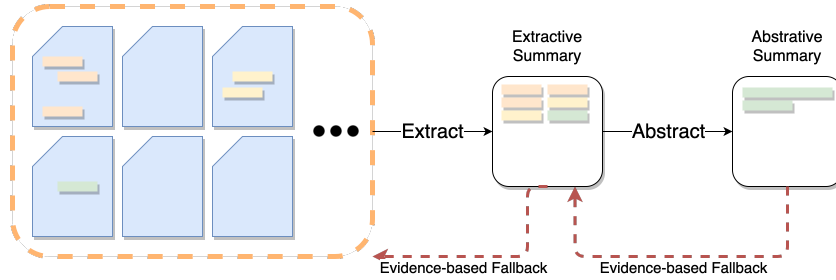


Figure 2: An extractive-abstractive summarization pipeline. The recall-oriented extractor extract relevant sentences from prior documents, the abstractor smooths out irrelevant or duplicated information.

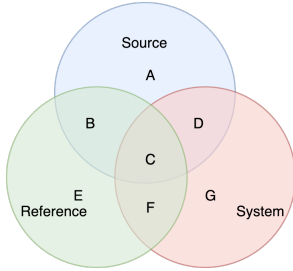


Figure 3: Relationship between source documents, reference summary, and system-generated summary.

Another advantage of this pipeline is that it provides a clear path of *evidence-based fallbacks* (Figure 2). Notably, both the extractor and the abstractor in the extractor-abstractor pipeline are capable of producing full summaries. Clinicians can reference the extractive summary if they find the abstractive summary problematic or if the abstractor model has low confidence. The extractive summary also has another level of fallback. The extracted sentences came from the source documents, so we can also display the extracted sentences in context or even use the extractor as a highlighter.

4 Measuring Faithfulness

Following prior work, we report ROUGE- n ($n = \{1, 2\}$) to measure n -gram overlap as a proxy for informativeness, and ROUGE-L (longest common subsequence, with possible gaps) as a proxy for fluency (Lin and Hovy, 2003; Maynez et al., 2020). However, as Schluter (2017) and Cohan and Goharian (2016) have argued, ROUGE alone is insufficient and possibly misleading for measuring informativeness, specifically when it comes to faithfulness and factualness.

In a summarization setting, a faithful summary refers to a summary that does not contain information from outside of the source. On the other hand, a factual summary allows information not presented in the source, as long as the information is factually correct. In the setting of clinical sum-

marization, we argue that faithfulness is far more important. Novel information appearing in a summary that has no support from the source, whether factual or not, can affect the transparency of the model.

A downside of this definition of faithfulness, however, is that it does not take reference summaries into account. Any extracted sentences (e.g., the first three sentences) from the source are always faithful by definition. Such extraction, however, might not be a summary relevant to this task. Figure 3 helps us illustrate the relationship between source documents, reference summary, and system-generated summary using a Venn diagram.¹

A desirable summary, especially in a clinical setting, is faithful to the source and relevant as measured by the reference summary. In Figure 3, this region corresponds to $B + C$, the ideal set of information a clinical summarization system should target. Based on this observation, we define *Faithfulness-adjusted Precision* as $\frac{C}{System}$ and *Faithfulness-adjusted Recall* as $\frac{C}{B+C}$. Intuitively, faithfulness-adjusted precision measures how much information in the system-generated summary is both relevant and faithful. Similarly, faithfulness-adjusted recall measures the amount of faithful and relevant information that has been included by the system. In a clinical setting, recall is often more important than precision; it is better to over-extract and have clinicians ignore or remove the irrelevant content than have missing content. While our extractive-abstractive pipeline provides a series of *fallbacks* that allows clinicians to inspect what could be missing by looking at the context of the extracted sentences, under-extraction can still hap-

¹Here we are showing a single reference summary, but in reality, the reference summary available is just one possible manifestation of all possible, potentially equally valid summaries (Nenkova and Passonneau, 2004). Our discussion can be extended to multiple reference summaries by treating each one independently in the calculations and averaging them to report the final scores.

pen. We therefore report a recall-oriented measure to combine the two above measures: *Faithfulness-adjusted F_β* , where we set $\beta = 3$. In this setting, faithfulness-adjusted recall is three times more important than faithfulness-adjusted precision (Van Rijsbergen, 1979).²

Hallucination is perhaps the leading concern of applying abstractive summarization system in a clinical setting. If one defines hallucination as a system generating content that is not faithful to the source, we can identify hallucination as the region $F + G$. G , the information that is not present in neither the source nor the reference, is particularly problematic. We therefore measure *Incorrect Hallucination Rate* as $\frac{G}{System}$.

However, an important underlying assumption of these measures is that the regions in Figure 3 are quantifiable. While there are many ways to approximate these regions, as a starting point, we use the medical named entity recognition (NER) system in SciSpacy (Neumann et al., 2019). The SciSpacy NER matches any spans in the text which might be an entity in UMLS, a large biomedical database, and transforms the text into a set of medical entities. The cardinalities of the sets and their overlaps can then be used to calculate the above measures.

5 Related Work

Clinical Summarization. Most literature on clinical summarization focuses on extractive summarization, due to the risk involved in a clinical application (Demner-Fushman and Lin, 2006; Feblowitz et al., 2011; Liang et al., 2019; Moen et al., 2016). For abstractive summarization, summarization of radiology reports has been a topic of interest in NLP research recently. Zhang et al. (2018) show promising results generating assessment section of a chest x-ray radiology report from the findings and background section. MacAvaney et al. (2019) improved this model through the incorporation of domain-specific ontologies. However, such generated reports may not be clinically sound, and the models generate sentences inconsistent with the patient’s background. Therefore, in subsequent work (Zhang et al., 2020) added a reinforcement learning based fact-checking mechanism to generate a clinically consistent assessment. Lee (2018) explores the generation of the *Chief Complaint* of emergency department cases from age

²We plan to explore the values of β in consultation with clinicians in future work.

| Dataset | Input | Output | # Data |
|---------------|-------------|--------------|--------|
| Gigaword | 10^1 | 10^1 | 10^6 |
| CNN/DailyMail | 10^2-10^3 | 10^1 | 10^5 |
| WikiSum | 10^2-10^6 | 10^1-10^3 | 10^6 |
| Our Dataset | 10^4-10^5 | $*10^0-10^3$ | 10^3 |

Table 1: Size comparison of summarization datasets. *For stats of the output sections, see Table 2.

group, gender, and discharge diagnosis code. Ive et al. (2020) follow a closely related approach of extracting keyphrases from mental health records to generate synthetic notes. They further evaluate the quality of generated synthetic data for downstream tasks. Work from Lee (2018) generates clinical notes by conditioning transformer-based models on a limited window of past patient data.

In our work, instead of focusing on purely extractive or abstractive clinical summarization, we proposed an extractive-abstractive pipeline as a framework for clinical multi-document summarization.

Faithfulness in Summarization. Recognizing the limitation of the existing measures and the danger of hallucination in summarization systems, faithfulness in summarization has gained attention recently (Kryscinski et al., 2020; Cao et al., 2017). Recent work on faithfulness evaluation in summarization involves using textual entailment (Maynez et al., 2020) or question answer generation (Aru-mae and Liu, 2019; Wang et al., 2020). For radiology summarization, Zhang et al. (2020) proposed using a radiology information extraction system to extract a pre-defined set of 14 pieces of factual information tailored to radiology reports.

In this paper, we approximate information overlap using the overlap of medical named entities. We argue that the domain of clinical encounter summarization is very different from the domains of most textual entailment tasks or question answer generation tasks. It is often much more specific, allowing us to apply the medical NER model. However, it is not as specific as the radiology summarization task, where a set of pre-defined information can more easily be identified.

6 Dataset

We derive our dataset from the MIMIC III database v1.4 (Johnson et al., 2016): a freely accessible, English-language, critical care database consisting of a set of de-identified, comprehensive clinical data of patients admitted to the Beth Israel Dea-

coness Medical Center’s Intensive Care Unit (ICU). The database includes structured data such as medications and laboratory results and unstructured data such as clinical notes written by medical professionals. For this work, we will focus on the unstructured data.

The challenge for adapting the MIMIC III database for our purpose, however, is that MIMIC III is incomplete. Due to the way that MIMIC III was collected, not all clinical notes are available; only notes from ICU, radiology, echo, ECG, and discharge summary (Johnson and Shivade, 2020) are guaranteed to be available. It is important to note that the incompleteness is not a property of the problem we are trying to address; it is a property of that database. We limit the incompleteness issue by focusing on the subset of encounters that contain at least one admission note (a clinical note written at the time of admission) as a proxy for completeness. This leaves us about 10% of the total encounters, or around 6,000 encounters.

We identify seven medical sections in the discharge summary as our targets for summarization: (1) chief complaint, (2) family history, (3) social history, (4) medications on admission, (5) past medical history, (6) history of present illness, and (7) brief hospital course. These medical sections were chosen based on their high prevalence in discharge summaries and their length diversity (see Table 2).

Target Section Extraction. To extract the target medical sections from the discharge summary, we use a regular expression based approach to identify the medical section headers’ variants from the training set. We then collect the content from the target medical section header and stop right before the next section header in the discharge summary. Around one hundred randomly selected extracted medical sections are manually examined to ensure no missing content or over-extraction. For each of these target medical sections, we then collect all the prior clinical notes (according to the chart date timestamp in MIMIC III) as their source documents. On average, the source documents consist of 64 documents and 36,3567 words. Table 1 shows a comparison with other dataset.

After the rule-based target extraction, we split the 6,000 encounters based on the *subject id* to prevent data leakage. Each section is split into training, validation, and test set (80/10/10) using the same set of subject ids. If the rule-based target extraction returns nothing, the encounter is excluded. See

Table 2 for the statistic of sample size.

7 Models and Experiments

As explained in Section 3, our proposed pipeline involves an extractive summarization component and an abstractive summarization component. This section identifies a set of existing extractors and abstractors across a diverse range of different approaches to understand what models are suitable for encounter-level clinical summarization. To understand the robustness of these approaches, we train and test these models across seven medical sections with a diverse range of length.

Extractors. Since our goal is to summarize an encounter conditioned on a targeted medical section, we focus our attention on supervised extractors. Supervised extractive summarization is often framed as a sentence extraction problem. Each sentence is encoded into a representation used to determine whether the sentence should be included in the extracted summary. RNN or transformer-based attention are often used to encode the surrounding sentences as context.

RNN+RL_{ext}: Chen and Bansal (2018) proposed a method to use reinforcement learning to fine-tune a pretrained RNN sentence extractor to a pointer network operating over sentences. By modeling the next sentence to extract (including the extra “end-of-extraction” sentence) as the action space, the current extracted sentences as the state space, and by using ROUGE between reference summary sentence and rewritten extracted sentence (rewritten by a separate pretrained abstractor) as the reward, the authors repurposed the sentence extractor to extract sentences from the source documents and reorder them as they might appear in the summary.

PRESUMM_{ext}: Liu and Lapata (2019b) proposed Presumm, a family of summarization models. Here we are especially interested in the extractive summarization variant that uses a modified pretrained BERT model to encode sentences to determine whether the sentence should be included in the extracted summary. While the model has been shown to achieve competitive results, applying a BERT encoder to very long text can be challenging in terms of memory limitations. Thus, we apply a split-map-reduce framework, where the long text is split into smaller units during training and inference. After inference, each smaller unit’s extracted sentences are then concatenated back together in

the same order as appeared in the original source. Since the model only assigns scores to sentences, we sweep the score cutoff threshold on the validation set using ROUGE-L score, and apply that cutoff on the test set.

Abstractors. In our extractive-abstractive pipeline, abstractors play a role in mapping the extracted sentences to the reference summary. Here we include two abstractor variants:³

RNN+RL_{abs}: Similar to RNN+RL_{ext}, however, after each sentence is extracted, it is immediately rewritten by passing through a pretrained sentence-level abstractor. The goal is to rewrite each extracted sentence to the format of what might appear in the reference summary. This sentence-rewriting approach has the disadvantage of only having a local view when rewriting (thus no merging of information). However, the advantage is that the memory limitation of sentence-level rewriting does not grow with the number of sentences, so it can be applied to longer summaries.

BART: Lewis et al. (2019) propose BART as a transformer variant that uses a bidirectional encoder similar to BERT and an autoregressive (left to right) decoder similar to GPT. The model has competitive performance for summarization, and thus is our choice for transformer-based abstractor. In contrast to the sentence-rewriting approach of RNN+RL_{abs}, we train BART to rewrite all the extracted sentences directly to the summary.

Baselines. Since clinical encounter summarization is a new task, there are no baselines from prior work. Following prior work on summarization, we include two special baselines: (1) ORACLE_{ext}: Extraction by using the reference summary; for each sentence in the reference summary, greedily select the source sentence in the source document that yields the maximum ROUGE-L score. (2) RULE-BASED_{ext}: apply the same rule-based target section extraction method in Section 6 that was used to construct the dataset. Instead of applying to the discharge summary, we apply the same extraction method to the prior clinical documents.

Evaluating the extractor-abstractor pipeline. For the two extractive models, RNN+RL_{ext} and PRESUMM_{ext}, as well as the two extraction baselines, we report ROUGE scores as well as our

³We also experimented with a pointer-generator (See et al., 2017), but we found that BART consistently outperforms pointer-generator, so we leave the results in the appendix.

proposed factualness-adjusted {precision/recall/F₃} scores across the seven medical sections.

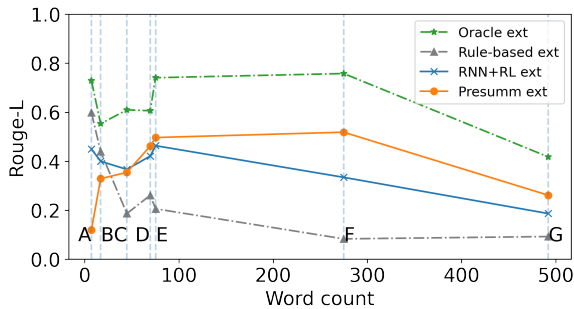
For the abstractive models, we measure the combinations of abstractive models with extractive models in our proposed pipeline. This implies measuring the performance of three models (two pointer-generator models shown in Appendix): RNN+RL_{abs} (uses RNN+RL_{ext} as the extractor), RNN+RL_{ext} + BART, and PRESUMM_{ext} + BART. For the abstractors, we additionally measure *incorrect hallucination rate* defined in Section 4.

8 Results and Discussion

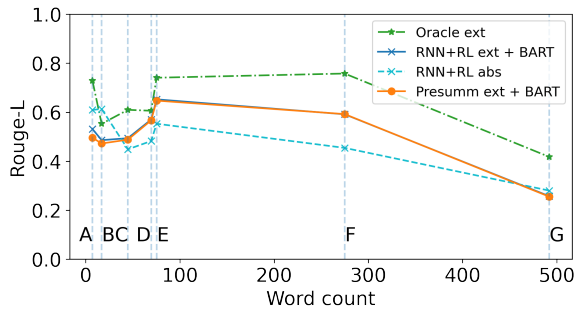
Extractive summarization and abstractive summarization are often applied in different settings and should thus be compared separately. For full results table, see Appendix A. Here we highlight the main findings. In Figure 4a, we highlight the ROUGE-L scores (ROUGE-1 and ROUGE-2 have a similar pattern) of the two extractive summarization systems compared to the oracle and rule-based extractive summary. An interesting observation is the effect of length: average word count of the reference summary medical section. RNN+RL_{ext} outperforms PRESUMM_{ext} on shorter sections, and vice-versa for the longer sections. This difference can be partially attributed to the way *cutoff* is being done at the extractors. For RNN+RL_{ext}, an RL agent is trained to decide when to stop extracting sentences. For the shorter sections, the RL learns to stop at just a few sentences (e.g., a typical chief complaint has two sentences, family history has on average 2.6 sentences). On longer sections, however, we find that the RL agent has difficulty stopping, causing over-extraction. In contrast, for PRESUMM_{ext}, a score cutoff threshold is tuned on the development set using the ROUGE-L score. This approach has a more balanced performance but suffers at short sections. Another factor contributing to the lead of PRESUMM_{ext} in the longer sections is our split-map-reduce framework, which enables the extractive model to conduct inference over all the clinical documents.

Interestingly, the baseline RULE-BASED_{ext} performs surprisingly well on Rouge for the two shortest sections. Upon inspection, most of the extraction is just the medical section’s title, without any content. This observation is backed up by the lower faithfulness-adjusted recall of this baseline.

For abstractive summarization, we highlight the ROUGE-L of the three abstractors in Figure 4b. Interestingly, after being abstracted by



(a) ROUGE-L of extractors vs. word count.



(b) ROUGE-L of abstractors vs. word count.

Figure 4: ROUGE-L of summarization models vs. average word lengths of the medical sections. Sections (dotted vertical lines) from short to long: (A) Chief complaint, (B) Family history, (C) Social history, (D) Medications on admission, (E) Past medical history, (F) History of present illness, and (G) Brief hospital course.

| | Chief Complaint | Family History | Social History | Medications on Admission | Past medical History | History of Present Illness | Brief Hospital Course |
|-------------------------------|-----------------|----------------|----------------|--------------------------|----------------------|----------------------------|-----------------------|
| train / val / test | 4,757/559/625 | 4,686/555/614 | 4,677/552/618 | 4,689/557/616 | 4,746/558/623 | 4,754/559/625 | 4,758/558/625 |
| Output # words | 7.25 | 17.03 | 44.90 | 69.58 | 75.36 | 274.88 | 491.97 |
| Output # sents | 2.04 | 2.63 | 4.93 | 4.67 | 5.99 | 16.62 | 35.39 |
| ORACLE _{ext} | 71.1/85.2/83.6 | 52.8/75.4/72.3 | 63.4/73.3/72.2 | 69.7/66.5/66.8 | 74.2/80.8/80.1 | 76.6/83.9/83.1 | 44.7/51.5/50.7 |
| RULE-BASED _{ext} | 97.4/49.7/52.2 | 87.6/47.3/49.6 | 94.7/23.1/25.0 | 97.2/32.8/35.2 | 94.9/16.9/18.4 | 70.8/08.6/09.5 | 00.3/00.9/00.7 |
| PRESUMM _{ext} | 10.8/24.1/21.4 | 30.7/63.1/57.1 | 42.6/40.6/40.8 | 48.7/52.0/51.7 | 51.2/66.6/64.7 | 54.4/74.5/71.9 | 26.5/47.7/44.2 |
| RNN+RL _{ext} | 44.2/72.8/68.4 | 54.5/70.6/68.6 | 43.2/71.0/66.7 | 45.7/67.2/64.2 | 43.6/81.7/75.1 | 27.6/88.8/72.7 | 15.3/69.7/51.4 |
| PRESUMM _{ext} + BART | 45.5/63.6/61.2 | 46.1/70.2/66.7 | 60.0/66.0/65.3 | 67.1/77.7/76.5 | 69.7/73.3/72.9 | 68.0/64.5/64.8 | 37.4/26.8/27.6 |
| RNN+RL _{ext} + BART | 48.6/70.4/67.4 | 44.7/74.2/69.6 | 61.2/66.7/66.1 | 67.0/80.2/78.7 | 70.0/74.6/74.2 | 67.4/64.7/64.9 | 34.1/23.6/24.4 |
| RNN+RL _{abs} | 67.8/69.1/69.0 | 75.8/73.0/73.3 | 60.1/68.2/67.3 | 70.9/69.0/69.2 | 64.7/68.8/68.3 | 40.8/82.2/74.6 | 20.4/52.9/45.6 |

Table 2: Dataset statistic and faithfulness-adjusted $\{Precision/Recall/F_3\}$ scores based on medical NER.

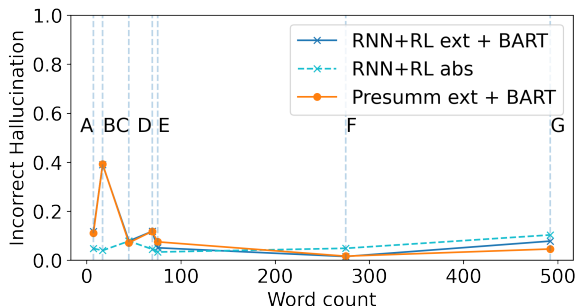


Figure 5: NER-based incorrect hallucination rate of abstractive models vs. average word lengths. Extractors do not hallucinate. Order the same as Figure 4.

BART, both RNN+RL_{ext} and PRESUMM_{ext} converged to roughly the same ROUGE-L scores. This suggests that in our extractor-abstractor pipeline, BART is effective in taking different extracted summaries and *smoothing* them into the format and content expected for the medical sections. On the other hand, RNN+RL_{abs} outperforms BART at the shorter sections, and even ORACLE_{ext} at the family history section. Note that ORACLE_{ext} is not necessarily an upper-bound for the abstractive summarization models; abstractors allow rewriting content in prior notes into the format of discharge

summary. Sentence segmentation (the basic unit of extraction) can also be noisy in clinical notes. On the other hand, the curve for RNN+RL_{abs} is almost identical to RNN+RL_{ext} in Figure 4a, with a constant increase. This is largely attributed to the sentence-rewriting of the sentence-level abstractor that allows RNN+RL_{abs} to keep the benefit of its extractor counterpart, while rewriting the content to reduce over-extracted sentences.

Table 2 shows our faithfulness adjusted measures. For the extractors, RNN+RL_{ext} outperforms all other extractors on faithfulness-adjusted F_3 and even outperforms ORACLE_{ext} in the brief hospital course section. This is possible because ORACLE_{ext} is selected using ROUGE-L, not faithfulness-adjusted F_3 . For the abstractors, a similarly good performance is found for RNN+RL_{abs}, where its precision consistently increases compared to RNN+RL_{ext}. The good performance of RNN+RL_{ext,abs} can largely be attributed to the high recall that has hurt their ROUGE-L performance in Figure 4. Interestingly, the two BART models again perform roughly the same, with recall of RNN+RL_{ext} + BART higher than PRESUMM_{ext} + BART. For the longest section, generation for

| | Summary |
|-------------------------------|--|
| ground_truth | past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3 # stable angina on long acting nitrate |
| Presumm _{ext} | # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3 nc occupation : changes to medical and family history : |
| RNN+RL _{ext} | # simvastatin 20 mg once a day # isosorbide mononitrate 40 mg once a day # furosemide 40 mg once a day # pantoprazole 40 mg once a day # diltiazem xr 180 mg once a day # tylenol for gum pain # proair hfa 90 mcg/actuation aerosol inhaler [hospital1] prn # prednisone per pt 's son 2 weeks ago # antibiotic for pneumonia per pt 's son 2 weeks ago past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3 nc occupation : sinus rhythm . |
| Presumm _{ext} + BART | past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3 |
| RNN+RL _{ext} + BART | past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.3 |
| RNN+RL _{abs} | past medical history : # hypertension # hyperlipidemia # gerd # ckd with baseline cr 1.5 . . |

Table 3: A random example showing summaries of past medical section. Despite RNN+RL_{ext} over-extracted in this example, BART was able to smooth out the noise and generate the same output Presumm_{ext} + BART.

BART proves to be difficult, as indicated by the large drop of recall, whereas the sentence-wise rewriting strategy of RNN+RL_{abs} has scaled better to longer sections.

The overall incorrect hallucination rate shown in Figure 5 is relatively low, with the notable exception of the family history section. Inspection of the generated summaries shows that the most common hallucination of both BART systems is the phrase “no family history”. Interestingly, the ground truths corresponding to these hallucinations are mostly variations of the term “non-contributory”; inspection of the source also shows that the family history section was often left blank. That being said, there are still cases of hallucinations where “no family history” is followed by a condition (e.g., arrhythmia, cardiomyopathies) that is not mentioned in the source.

Table 3 shows a qualitative analysis of a randomly chosen summary of the past medical section. In this case, RNN+RL_{ext} over-extracted content from the previous sections. However, after passing through BART, BART successfully smooths out the noise and generates the same output as Presumm_{ext} + BART. In this case, RNN+RL_{abs} happens to be hallucinating (mapping cr 1.3 to cr 1.5). All summarization systems missed “# stable angina on long acting nitrate”; mention of “angina” is actually not present in the prior clinical notes.

9 Conclusion and Future Work

We present a novel clinical summarization task – discharge summary composition by summarizing prior clinical documents, derived from a public database (MIMIC III). By summarizing the vast number of clinical notes in a format clinicians are

already trained to read and understand, our work has the potential to reduce the time clinicians spend on making sense of the data, allowing them to allocate more time to the patients.

We view this work as a promising first step to measure how existing models perform on the task and share the task with the community. One limitation of this work is that if there is novel information available only when writing the discharge summary, there will be no way of summarizing it. It is also important to note that since we are using MIMIC III for training and evaluation, the results shown are biased to the dataset, as MIMIC III is an English-language collection from the ICU of a single hospital, and may not necessary be applicable to other clinical setting.

We identify three main challenges: (1) faithfulness, (2) evidence, and (3) long text. An extractor-abstractor pipeline is proposed to provide a natural way of fallback with an increasing amount of evidence at each fallback and also enable scaling to very long documents. To investigate the risk of hallucination and faithfulness in the summaries, we evaluate with a NER-based measure on top of ROUGE. Adapting state-of-the-art summarization models, our experiments over seven medical sections demonstrate the potential for the extractor-abstractor pipeline and represent a framework towards a set of enabling technologies that can assist clinicians to better make sense of the vast amount of unstructured data in the EHR.

10 Ethical Considerations

Deidentification. Our dataset is derived from the public database MIMIC III v1.4 (Johnson et al.,

2016). Johnson et al. (2016) deidentified the database in accordance with the Health Insurance Portability and Accountability Act (HIPAA) standards. This standard requires removing all eighteen identifying data elements, including patient name, telephone number, address, and dates. These fields are replaced with placeholders. A constant (but different per patient) offset is applied to shift the dates. Patients over 89 years old were mapped to over 300, in compliance with HIPAA.

Although under U.S. federal guidelines, secondary use of fully deidentified, publicly available data is exempt from institutional review board (IRB) review (45 CFR § 46.104, “Exempt research”), we still consider the dataset sensitive. We are careful to treat it as such. During training and error analysis, we of course do not attempt to identify individuals, and when the qualitative analysis is shown, we double-check to avoid showing potentially identifiable information.

Population. In MIMIC III, out of the 38,161 patients, 71.34% are White, 7.69% Black, 2.38% other, 2.37% Asian, and the rest unknown. Most of the patients in MIMIC III were older adults, with the most common age group being 71–80, followed by the 61–70 age group. (Dai et al., 2020).

Broader Impact. Clinical application has the genuine potential to affect people’s lives. As we have emphasized in Section 1 and Section 9, this work is not about a discussion for deployment, but rather a first step in understanding how the current existing summarization models perform. Importantly, we need to understand the failure modes of these systems and how to address these failures. Our emphasis on faithfulness and traceability of summarization reflects those beliefs. Hopefully, the three challenges we identify are the first of many future steps that will make progress toward alleviating the documentation burden of clinicians and ultimately result in a better quality of care for patients.

References

Kristjan Arumae and Fei Liu. 2019. Guiding extractive summarization with question-answering rewards. 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 2566–2577.

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

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the original: Fact aware neural abstractive summarization. *arXiv preprint arXiv:1711.04434*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of ACL*, pages 675–686.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 806–813.

Zheng Dai, Siru Liu, Jinfa Wu, Mengdie Li, Jialin Liu, and Ke Li. 2020. Analysis of adult disease characteristics and mortality on mimic-iii. *PLOS ONE*, 15(4):1–12.

Dina Demner-Fushman and Jimmy Lin. 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 841–848.

Joshua C. Feblowitz, Adam Wright, Hardeep Singh, Lipika Samal, and Dean F. Sittig. 2011. Summarization of clinical information: A conceptual model. *Journal of Biomedical Informatics*, 44(4):688 – 699.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for Natural Language Processing. *npj Digital Medicine*, 3(1):1–9.

Alistair Johnson and Chaitanya Shivade. 2020. Notes and data not in mimic-iii · issue 771 · mit-lcp/mimic-code.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

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.

Scott H. Lee. 2018. Natural language generation for electronic health records. *npj Digital Medicine*, 1(1):1–7.

- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- 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.
- Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.
- Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.
- Yang Liu and Mirella Lapata. 2019a. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081.
- Yang Liu and Mirella Lapata. 2019b. Text summarization with pretrained encoders. In *Proceedings of EMNLP*, pages 3721–3731.
- Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W. Filice. 2019. [Ontology-Aware Clinical Abstractive Summarization](#). In *Proceedings of SIGIR*, pages 1013–1016. Association for Computing Machinery, Inc.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
- 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.
- Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.
- Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. [ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.
- Natalie Schluter. 2017. The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45.
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of ACL*, pages 1073–1083.
- Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. [Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties](#). *Annals of Internal Medicine*, 165(11):753–760. PMID: 27595430.
- Daniel S Tawfik, Jochen Profit, Timothy I Morgenthaler, Daniel V Satele, Christine A Sinsky, Liselotte N Dyrbye, Michael A Tutty, Colin P West, and Tait D Shanafelt. 2018. Physician burnout, well-being, and work unit safety grades in relationship to reported medical errors. In *Mayo Clinic Proceedings*, volume 93, pages 1571–1580. Elsevier.
- Cornelis J Van Rijsbergen. 1979. Information retrieval. (2nd ed.). *University of Glasgow*, pages 133–134.
- Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. [Asking and answering questions to evaluate the factual consistency of summaries](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.
- Michael Weiner and Paul Biondich. 2006. The influence of information technology on patient-physician relationships. *Journal of general internal medicine*, 21(1):35–39.
- C. P. West, L. N. Dyrbye, and T. D. Shanafelt. 2018. [Physician burnout: contributors, consequences and solutions](#). *Journal of Internal Medicine*, 283(6):516–529.
- Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. [Big bird: Transformers for longer sequences](#).
- Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. [Learning to Summarize Radiology Findings](#). In *Proceedings of the Workshop on Health Text Mining and Information Analysis (EMNLP-LOUHI)*, pages 204–213. Association for Computational Linguistics (ACL).

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.

A Appendix: Full Results

See Table 4 and Table 5 for the full scores for all models on all seven sections.

B Appendix: Qualitative Analysis

Table 6 shows a random example on social history section.

C Appendix: Reproducibility

Here we describe the training details of the models for reproducibility.

RNN+NL_{ext} and RNN+NL_{abs}. Both models are trained following the original recipe (Chen and Bansal, 2018). The training setup involves the following steps: (1) use gensim to train a word2vec embedding from scratch from the training set of the source documents, (2) construct pseudo pairs of sentences (source sentence, summary sentence): for each summary sentence, greedily finds the one-best source sentence using ROUGE-L recall, (3) use the pseudo pairs to train an RNN extractor, (4) use the pseudo pairs to train a pointer-generator that rewrites the sentences, and (5) train an RL agent that fine-tunes the RNN extractor with the sentence-rewriting pointer-generator. Model is trained on one V100 GPU, with an Adam optimizer of learning rate 1e-3. Here we use the same set of hyperparameters as Chen and Bansal (2018). For more details, please refer to the original paper.

For each of the seven medical sections, we follow the training recipe, and repeat it five times. The reported models are chosen based on the validation set. We found that the RL fine-tuning step can potentially be very unstable. For the longer sections (e.g., brief hospital course and history of present illness), the RL fine-tuning can even fail to converge.

PRESUMM_{ext}. We use the original implementation released with PRESUMM (Liu and Lapata, 2019b). Learning rate is set to 2e-3 and extractor dropout rate is set to 0.1, following the original

paper. `bert-base-uncased` is used as the pre-trained BERT model. We made three important changes: (1) increase the maximum tokens the encoder can consume to 1024 tokens, (2) in the data preprocessing step, we construct pseudo pairs of sentences that will be later used to train the extractor: for each summary sentence, greedily finds the one-best source sentence using ROUGE-L recall, and (3) before the training begin, we split the source documents and their labels into segments smaller than 1024 tokens. After inference finishes, we concatenate the segments (together with a extraction score for each sentence) back together in the original order.

For each of the seven medical sections, we train the model on 4 V100 GPUs, with 150,000 training steps and model checkpointing every 2,000 steps. We report the model with the lowest model loss on the validation set. Since the model only assigns scores to sentences, we sweep the threshold of score cutoff on the validation set using ROUGE-L score, and apply that cutoff on the test set.

POINTGEN. We use an open implementation of pointer-generator (See et al., 2017), implemented with PyTorch and AllenNlp.⁴ Our model follows the original paper and has 256-dimensional hidden states and 128-dimensional word embeddings. The vocabulary size is set to 50k words for both source and target. The model is optimized using Adagrad with learning rate 0.15 and an initial accumulator value of 0.1, and trained on one v100 GPU for 50 epochs with early stopping on the validation set.

BART. We use the Fairseq (Ott et al., 2019) implementation of BART-large (Lewis et al., 2019) as it is shown to achieve the state-of-the-art ROUGE scores for abstractive summarization. We fine-tune the BART-large model with the standard learning rate of 3×10^{-5} . We utilize a machine with 8 GPUs and batch size of 2048 input tokens per GPU. We train for a maximum of 10 epochs with early stopping to select the checkpoint with the smallest loss on the validation set. During decoding, we use beam search with beam size of 6. We restrict the generation length to be between 10 to 300 tokens.

⁴<https://github.com/kukrishna/pointer-generator-pytorch-allennlp>

| | Chief Complaint | Family History | Social History | Medications on Admission | Past medical History | History of Present Illness | Brief Hospital Course |
|------------------------|-----------------------|-----------------------|-----------------------|--------------------------|-----------------------|----------------------------|-----------------------|
| Oracle ext | 73.0/59.0/72.9 | 55.7/40.5/55.3 | 62.0/48.2/61.0 | 61.5/47.7/60.6 | 75.1/67.0/74.1 | 77.4/66.8/75.8 | 45.7/22.3/41.8 |
| Rule-based ext | 59.8/44.5/59.8 | 43.9/31.8/43.9 | 18.6/12.1/18.6 | 26.1/22.2/26.1 | 20.6/16.3/20.6 | 8.3/7.3/8.3 | 9.2/8.5/9.2 |
| RNN+RL ext | 45.1/33.1/45.0 | 40.2/28.6/40.0 | 37.6/27.2/36.6 | 43.4/35.6/42.1 | 47.9/40.2/46.3 | 34.8/28.3/33.4 | 21.3/6.7/18.6 |
| Presumm ext | 12.3/6.9/11.9 | 33.2/24.0/32.9 | 36.3/27.5/35.4 | 47.2/40.7/46.2 | 50.8/41.9/49.7 | 53.2/45.4/51.8 | 29.6/10.6/26.1 |
| RNN+RL ext + PointGen | 21.2/13.2/21.1 | 29.8/22.0/29.5 | 36.7/26.3/36.2 | 49.2/41.7/48.1 | 46.3/38.6/45.0 | 38.8/28.3/37.4 | 20.6/8.6/19.2 |
| Presumm ext + PointGen | 19.8/11.6/19.7 | 30.6/23.5/30.5 | 42.5/31.1/41.4 | 50.0/43.0/49.0 | 52.4/45.0/51.2 | 43.0/35.2/41.6 | 20.9/9.6/19.4 |
| RNN+RL ext + BART | 53.5/37.5/53.1 | 48.9/38.6/48.6 | 50.3/38.0/49.4 | 58.2/51.9/57.0 | 66.9/58.5/65.2 | 61.1/51.3/59.1 | 28.2/10.6/25.7 |
| Presumm ext + BART | 49.9/33.0/49.6 | 47.4/37.5/47.2 | 49.6/38.3/48.8 | 57.8/50.9/56.7 | 66.0/58.3/64.7 | 61.0/52.4/59.2 | 28.0/12.4/25.5 |
| RNN+RL abs | 61.2/47.5/60.9 | 61.6/50.5/61.3 | 45.9/33.7/44.8 | 49.9/42.2/48.2 | 57.5/47.9/55.3 | 47.6/38.4/45.4 | 32.1/10.4/28.0 |
| # words | 7.25037 | 17.026 | 44.9034 | 69.5803 | 75.3616 | 274.881 | 491.971 |
| # sents | 2.04183 | 2.63082 | 4.92901 | 4.67285 | 5.99115 | 16.6193 | 35.389 |

Table 4: ROUGE- $\{1/2/L\}$ scores, across different models and sections

| | Chief Complaint | Family History | Social History | Medications on Admission | Past medical History | History of Present Illness | Brief Hospital Course |
|-----------------------------------|-----------------|----------------|----------------|--------------------------|----------------------|----------------------------|-----------------------|
| ORACLE _{ext} | 71.1/85.2/83.6 | 52.8/75.4/72.3 | 63.4/73.3/72.2 | 69.7/66.5/66.8 | 74.2/80.8/80.1 | 76.6/83.9/83.1 | 44.7/51.5/50.7 |
| RULE-BASED _{ext} | 97.4/49.7/52.2 | 87.6/47.3/49.6 | 94.7/23.1/25.0 | 97.2/32.8/35.2 | 94.9/16.9/18.4 | 70.8/08.6/09.5 | 00.3/00.9/00.7 |
| PRESUMM _{ext} | 10.8/24.1/21.4 | 30.7/63.1/57.1 | 42.6/40.6/40.8 | 48.7/52.0/51.7 | 51.2/66.6/64.7 | 54.4/74.5/71.9 | 26.5/47.7/44.2 |
| RNN+RL _{ext} | 44.2/72.8/68.4 | 54.5/70.6/68.6 | 43.2/71.0/66.7 | 45.7/67.2/64.2 | 43.6/81.7/75.1 | 27.6/88.8/72.7 | 15.3/69.7/51.4 |
| RNN+RL _{ext} + POINTGEN | 40.6/70.2/65.4 | 38.2/73.9/67.6 | 59.9/58.7/58.8 | 66.4/72.7/72.0 | 65.6/59.0/59.6 | 69.1/37.1/38.9 | 39.8/15.2/16.2 |
| PRESUMM _{ext} + POINTGEN | 31.3/62.6/56.9 | 37.0/72.3/66.0 | 54.7/61.9/61.1 | 65.1/73.7/72.8 | 64.0/62.6/62.7 | 69.8/42.4/44.1 | 42.2/17.9/19.0 |
| PRESUMM _{ext} + BART | 45.5/63.6/61.2 | 46.1/70.2/66.7 | 60.0/66.0/65.3 | 67.1/77.7/76.5 | 69.7/73.3/72.9 | 68.0/64.5/64.8 | 37.4/26.8/27.6 |
| RNN+RL _{ext} + BART | 48.6/70.4/67.4 | 44.7/74.2/69.6 | 61.2/66.7/66.1 | 67.0/80.2/78.7 | 70.0/74.6/74.2 | 67.4/64.7/64.9 | 34.1/23.6/24.4 |
| RNN+RL _{abs} | 67.8/69.1/69.0 | 75.8/73.0/73.3 | 60.1/68.2/67.3 | 70.9/69.0/69.2 | 64.7/68.8/68.3 | 40.8/82.2/74.6 | 20.4/52.9/45.6 |

Table 5: Faithfulness-adjusted $\{Precision/Recall/F_3\}$ scores based on medical NER.

| | Summary |
|-------------------------------|--|
| ground_truth | social history : retired from [country 11150] . brother and son are part of support network . |
| Presumm _{ext} | [last name (un) 574] : retired gentleman from [country 4952] ; currently living with sons who are his main caretakers . . pt is hindi speaking only but able to communicate his needs and pleasant and cooperative . |
| RNN+RL _{ext} | family / social history : retired gentleman from [country 4952] ; currently living with sons who are his main caretakers . . saw pt ; did carotid massage ; give lopressor 50 mg po bid starting tonight social history : |
| Presumm _{ext} + BART | social history : [last name (un)] : retired gentleman from [country] ; currently living with sons who are his main caretakers . pt is hindi speaking only but able to communicate his needs and pleasant and cooperative . |
| RNN+RL _{ext} + BART | social history : retired gentleman from [country 651] ; currently living with sons who are his main caretakers . |
| RNN+RL _{abs} | social history : retired gentleman from [country]] ; currently living with sons who are his main caretakers . |

Table 6: A random example showing summaries of social history section.