Annotate the Way You Think: An Incremental Note Generation Framework for the Summarization of Medical Conversations

Longxiang Zhang, Caleb Hart, Susanne Burger, Thomas Schaaf
Solventum (3M Health Information Systems)
Pittsburgh, PA, USA
{lzhang28, chart2, smburger, tschaaf}@solventum.com

Abstract

The scarcity of public datasets for the summarization of medical conversations has been a limiting factor for advancing NLP research in the healthcare domain, and the structure of the existing data is largely limited to the simple format of conversation-summary pairs. We therefore propose a novel Incremental Note Generation (ING) annotation framework capable of greatly enriching summarization datasets in the healthcare domain and beyond. Our framework is designed to capture the human summarization process via an annotation task by instructing the annotators to first incrementally create a draft note as they accumulate information through a conversation transcript (Generation) and then polish the draft note into a reference note (Rewriting). The annotation results include both the reference note and a comprehensive editing history of the draft note in tabular format. Our pilot study on the task of SOAP note generation showed reasonable consistency between four expert annotators, established a solid baseline for quantitative targets of inter-rater agreement, and demonstrated the ING framework as an improvement over the traditional annotation process for future modeling of summarization.

Keywords: Annotation, Summarization, Incremental Generation, Medical Conversations

1. Introduction

Automatic clinical documentation, in particular the documentation of **Do**ctor-**Pa**tient **Co**nversation (DoPaCo), has been a focal point of NLP research in recent years (Ben Abacha et al., 2023a,b; Sharma et al., 2023). The practical impact is clear: having an automatic system that summarizes medical reports based on DoPaCos helps liberate doctors/physicians from the burden of documentation, a major factor contributing to physician burn-out (West et al., 2018). Even in scenarios where human assistants (known as medical scribes) are assigned to document the conversations, an automatic summarization model can provide value by assisting the scribes in improving efficiency and accuracy.

However, the landscape of summarization datasets in the healthcare domain limits current NLP research on clinical documentation in several aspects: first, although public datasets on general dialogue summarization exist in relative abundance (e.g., AMI corpus (Carletta, 2006), SAMSum (Gliwa et al., 2019), and DialogSum (Chen et al., 2021)), DoPaCo datasets for summarization are scarce and generally small. PriMock57 (Papadopoulos Korfiatis et al., 2022) is a high-quality summarization dataset of mocked consultation visits with 57 conversations; ACI-Bench (Yim et al., 2023) is the largest public dataset for DoPaCo summarization with 207 conversations. Second, almost all existing datasets on summarization rely on a "transcript-summary" pair format where one complete transcript is associated with one or more complete reference summaries. This format is only suitable for training deep learning models in an end-to-end (E2E) fashion, where the model consumes a complete transcript and outputs a complete summary; therefore, the trained models are effectively blackboxes with limited interpretability. Within and outside of the medical domain, people have been trying to break the limit of E2E training by introducing multi-stage training paradigms (Krishna et al., 2021; Zhang et al., 2021; Su et al., 2022), constructing intermediate output (Wang et al., 2022), or injecting knowledge graphs into the model input (see Gao et al. (2023) and the references therein). However, the intermediate data required for training the added components in these methods are either inferred from the transcript-summary pair or annotated by the researchers for specific purposes.

Furthermore, the format of transcript-summary pairs only represents the final output after an annotation process (summarization), not the process itself. Summarization is typically a cumulative process for humans: the annotators keep drafts or partial key notes, build a connection between evidences in the conversation and items in the notes, and revise or remove content based on new information as they read through a dialogue transcript (see for example, Knoll et al. (2022)). None of these intermediate steps are captured in the existing DoPaCo datasets, and limited research has been done on an annotation framework or tool appropriate for incorporating these steps: He et al. (2022) demoed MedTator as a general purpose annotation tool that could capture enriched intermediate data, but the tool was designed for clinical

LREC-COLING 2024, pages 1173–1186

documents and no study was done on its fitness for dialogue summarization. Perry et al. (2020) proposed a refined annotation schema for clinical conversations, but their goal was to extract clinical concepts, not to create clinical notes from the conversations. Yim et al. (2020) explicitly outlined an annotation schema for aligning and grouping evidences in a clinical visit with individual items in a clinical note; however, their proposed schema was based on finished notes and the linked evidence data was inferred *a posteriori*.

In this paper, we introduce a novel annotation framework we call Incremental Note Generation (ING) that aims to emulate the human process of clinical documentation (Figure 1). The design principle of the task is to capture all necessary data that reflects the cumulative nature of a human summarizing a conversation in real time, thereby resulting in annotated data of rich content and structure. In addition to the traditional transcript-summary pair data, the annotation results from ING evidently include (a) mapped evidences for every sentence in the final note; (b) a series of temporally ordered partial notes that align with incremental views of the conversation; (c) a complete and structured revision history between any two adjacent partial notes. We propose that our ING framework provides the necessary annotation to facilitate multi-task training of deep learning models with a diverse selection of objectives. Although the ING task is designed for clinical documentation, it can be easily adopted for any other summarization task and enrich the structure of the resulting dataset. To the best of the authors' knowledge, this is a first attempt in formulating the human thought process of summarization as an annotation task.

2. Task Definition

We propose the annotation framework for Incremental Note Generation (ING) as illustrated in Figure 1. This task is inspired by the workflow of synchronous scribing, where medical scribes are expected to complete a clinical note parallel to a live conversation. Because we are focusing on the healthcare domain, the annotators are expected to have medical expertise and experience in generating clinical notes. The ING framework consists of two main sequential steps: Generation and Rewriting.

2.1. Generation

Draft note generation, or simply Generation (left block in Figure 1), is the first and foremost component in a complete ING workflow. The annotator is required to read through a DoPaCo transcript sequentially and construct a draft note in an incremental way. To ensure the "incremental" nature of the annotation process, we explicitly forbid annotators to look ahead in the transcript. In addition, we allow three types of actions during the annotation:

ADD When sufficient medical information has been discussed at any point in the conversation, mark ALL supporting information (hereafter referred to as "evidences") and create a note item to record the information. Multiple items in different sections can be created from the same evidence(s).

UPDATE When new information appears that changes, modifies, or expands any of the existing note item(s), mark the evidences that support the change and update the relevant item(s). Multiple items can be updated and merged into a new item.

REMOVE When new information appears that invalidates a note item, remove the note item and mark the evidences that support the removal.

As the annotator finishes a conversation transcript, the complete history of the annotator's decision is recorded and all medically relevant information is added/updated/removed into a list of note items, which we term "draft note".

2.2. Rewriting

The draft note resulting from Section 2.1 may not be "deliverable" because of unpolished language and unrefined format. To refine the draft note into a standard clinical note we add the Rewriting task (top right block in Figure 1). Although Rewriting depends on the Generation for input, we intend to keep the two tasks independent otherwise. Specifically, we require annotators to NOT consult the original conversation during rewriting, thereby restricting their edits to be purely linguistic and information preserving. Common edits such as reordering, merging, or splitting of note items are supported (encouraged actually) as long as those edits do not add, remove, or change the meaning of affected note items. As long as enough draft notes are available from the generation task, Rewriting can be conducted by a different group of annotators in parallel.

2.3. Evidence Mark-Up

An optional third task (bottom right block in Figure 1), named Evidence Mark-Up (EMU) can be incorporated in parallel with the Rewriting task. This is to delegate the burden of marking evidences from medical experts to non-experts and help distribute the workload for improved efficiency, and shares certain similarity with Yim et al. (2020). It complements the Generation task and is outlined as follows:

Mark boundaries In the generation task, annotators are no longer required to mark all evidences for every action they take but mark only the last evidence or a "boundary" in the conversation at

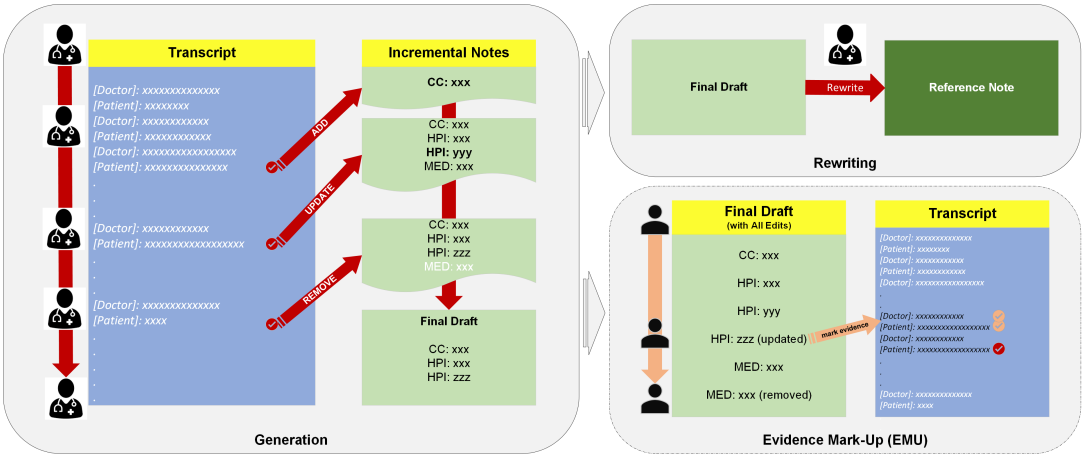

Figure 1: The Incremental Note Generation (ING) Framework: **Generation:** Annotator with medical expertise reads through Transcript sequentially, marks evidences along the way, creates Final Draft incrementally by adding, updating, or removing items. **Rewriting:** Annotator with medical expertise reads through Final Draft and rewrites it into professional clinical note without referring Transcript. **Evidence Mark-Up (EMU):** (optional) Annotator with/without medical expertise reads through Final Daft with the editing history, and searches for evidences in Transcript. Note that the last marked evidence for each note item from the Generation step (red checkmark) is provided to bound the searching region.

which sufficient information is present to support their ADD/UPDATE/REMOVE actions.

Recover marked evidences A (different, possibly non-expert) annotator is presented with the transcript, the draft note, and the marked boundaries, and is required to find and mark all evidences for each note item within the specified boundaries.

Mark only the changes Draft notes presented to EMU annotators should contain all note items created during the Generation task, including the ones updated and removed. For those items, the annotators are tasked with finding evidences that support the change.

Subjective	CC - Chief complaint HPI - History of present illness ROS - Review of systems PMH - Past medical history SOCH - Social history FAMH - Family history ALGY - Allergy MED - Medication
Objective	Vital Signs Treatment/Procedures Imaging & Labs PE - Physical exams
Assessment/Plan	Assessment Plan Follow-up

Table 1: Medical sections used in the ING task.

3. Guidelines and GUI

In this section, we choose the specific task of **SOAP** (Subjective, Objective, Assessment, and Plan) note generation(Podder et al., 2022) from DoPaCos to ground our discussion on the guidelines and the design of the annotation GUI.

3.1. Guidelines

The guidelines for ING were created collaboratively with a group of three linguists (one with medical background), five medical scribes (at least six months of experience), and one practicing medical doctor as the advisor. The team conducted multiple rounds of trial annotations on three example conversations and then iteratively updated the guidelines through a series of feedback meetings. Since our target annotators are professional medical scribes, the guidelines focus on an overview of the ING framework, a tutorial of the GUI, and a list of formatting requirements such as "use a concise sentence for each added note item except when updating items" and "copy the text verbatim if it comes from a doctor's dictation". The content selection is delegated to the medical expertise of individual scribes. The only specific requirement regarding the structure of the SOAP note is represented in Table 1, which shows an exhaustive list of all medical sections that annotated note items should be classified into[1]. Detailed guidelines are included in Appendix A.

3.2. GUI

The GUI for the Generation task (Section 2.1) is developed in Python and is shown in Figure 2. The

[1]The list of sections is also refined iteratively to be representative of most common use cases.

most outstanding components (indexed as in the figure) are: **(1)** The navigation panel.[2] **(2)** The transcript table. This scroll-able table supports common interactive features such as changing the column width, auto-wrapping of overflowing text, and multi-row selection/deselection. **(3)** The annotation tools. We implement a drop-down list with all section headers from Table 1, an empty text box for text editing, an "Action Status" button (see (6)), Undo/Redo buttons for correcting mis-operations, two "Clear ..." buttons for quick reset of selected rows in either the transcript or the note item table, and a "Clear current note" button for restarting the work from scratch. **(4)** The note item table. This table contains three tabs (Subjective, Objective, Assessment/Plan), each displays all current note items of one category. Updated or removed items due to UPDATE/REMOVE actions are hidden. Users can sort by column, the default sorts by the order of the generation of all note items. **(5)** The draft note preview. This is a non-editable text box that displays the current state of the note derived from the table above. **(6)** An Action Status button that adds the current annotation to component (4) only when the state of the GUI meets certain criteria.

The **Action Status Button** was added after the annotators asked for a more streamlined and directed design to clarify the complex process and reduce the number of steps and their order to be memorized. It only becomes clickable and changes the label into one of **ADD/UPDATE/REMOVE/EDIT**, when the GUI is in a "valid" state, i.e. when settings in **(2)**, **(3)**, and **(4)** reflect the intention of each action (an example of a valid UPDATE is shown in Figure 2). Table 2 lists in detail the valid GUI conditions for each action. The **EDIT** action is included for annotators to correct simple typos or an incorrect section as an alternative to Undo/Redo buttons. Through this button, the definition of each action in Section 2.1 is embedded as visual features, so that annotators can understand the ING framework organically through the use of the GUI.

The GUI for the EMU step is largely identical to Figure 2. A detailed layout is presented in Appendix B. We defer our discussion on a GUI for the Rewriting step to Section 5.

3.3. Output

Table 3 shows the data structure of ING annotations, specifically the expected output in the Subjective category of a SOAP note from either the Generation step (Section 2.1) or from Generation + EMU if an EMU step (Section 2.3) is adopted. The tabular structure with the given columns offers great flexibility in generating a wide range of derived data. First, a final draft note can be extracted from the rows with $show\_flag = True$ with additional grouping by section and sorted by the marked evidences in the $rows$ column (as is done in component (5) in Figure 2). Second, a series of incremental partial notes can be created by accumulating row data in the table, and the corresponding incremental views of the transcript can be determined by the last marked evidence in each partial note. Third, an ordered series of actions or editing history can be derived based on the values in *next_rows* and *removed* columns; for example, the second and third rows in Table 3 indicate the third row is updated from the second row, which by itself is added; the last row indicates this note item is added from lines 138-139 in the transcript but then removed based on lines 201, 202, and 205.

GUI state variables	
N_{conv}: No. of selected transcript rows (component (2))	
N_{note}: No. of selected note items (component (4))	
B_{sec}: If a section is selected (component (3))	
B_{item}: If the text box is non-empty (component (3))	
Inferred action	**Conditions**
ADD	$(\geq 1, 0, \text{True}, \text{True})$
UPDATE	$(\geq 1, \geq 1, \text{True}, \text{True})$
REMOVE	$(\geq 1, 1, -, \text{False})$
EDIT	$(0, 1, \text{True}, \text{True})$

Table 2: Conditions for the Action Status button in the GUI (Figure 2). Each condition is represented by a tuple of four state variables: $(N_{conv}, N_{note}, B_{sec}, B_{item})$ with definition listed at the top of the table. $-$ means the corresponding state variable is ignored.

4. Pilot Run & Data Analysis

We conducted a pilot run of the ING task with all three steps (Section 2.1 - 2.3) on 23 DoPaCos. Three conversations were used to onboard annotators for the task (referred hereafter as "training phase") and actual annotations were done on 20 conversations in the family medicine specialty (referred hereafter as "pilot phase"). Four experienced medical scribes were hired as annotators for both the Generation and Rewriting steps, while five linguists with no medical expertise were employed as the annotators for the EMU step.

During the training phase, each scribe was given an overview of the task, a tutorial on the GUI, and then requested to finish the Generation step on the three conversations; feedback sessions were then organized between all scribes and a domain expert. During those sessions, the domain expert and the authors corrected possible misunderstanding or mistakes made in the draft notes, clarified disam-

[2]The panel should be self-explanatory: the "Back to Login Page" brings the user back to a standard login page with user name and password input which we choose not to show due to space limit.

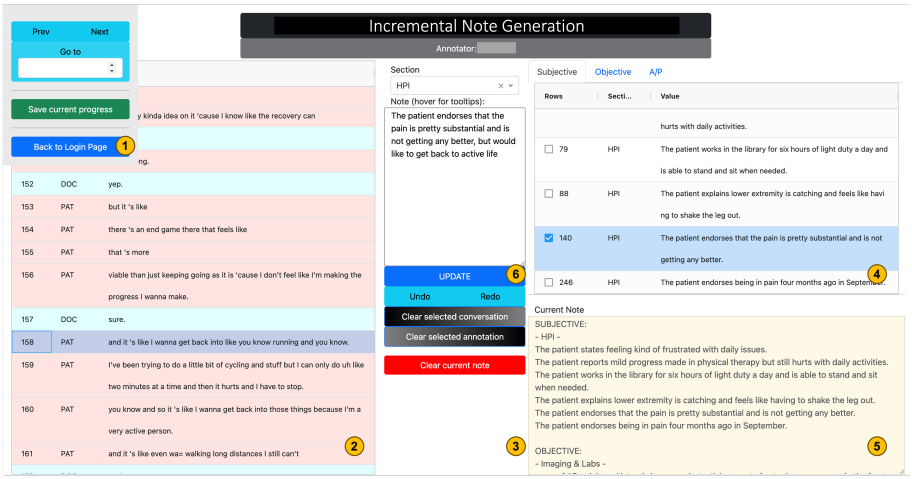

Figure 2: GUI for ING Generation task. Indexed components are: (1) navigation panel; (2) transcript table; (3) annotation tools; (4) note item table; (5) draft note preview; (6) action status button.

rows	next_rows	section	value	removed	show_flag
49, 50, 51	None	HPI	The patient is currently taking Prednisone.	False	True
42, 43	53, 54	CC	The patient complains of pain.	False	False
53, 54	None	CC	The patient complains of shoulder pain.	False	True
63, 66	None	HPI	The patient left knee is swelling.	False	True
138, 139	201, 202, 205*	HPI	The patient is scheduled for an MRI on Monday.	True*	False*

Table 3: Example ING output data structure. The table shows a portion of the Subjective category. *: The values are artificially changed to showcase an example annotation from a REMOVE action.

biguities on the guidelines, and collected scribes' feedback on the task, which directly contributed to the iterative updates of our guidelines and GUI. The pilot phase followed afterwards and scribes were instructed to independently finish the remaining 20 DoPaCos. An EMU pass was then conducted on the 80 (20×4) draft notes by the linguists with minimal training (a demo session on one example conversation). We leveraged a mix-and-match strategy to ensure that each EMU annotator was assigned draft notes created by all four scribes and each draft note was marked by more than one annotator. All scribes were further instructed to finish the Rewriting step on six of the 20 draft notes they created.

4.1. Basic statistics

The 20 DoPaCos contain on average 300 (std. 140) lines of conversation with an average count of 1667 words (std. 722). Through self-reporting, we estimate the average speed for one scribe to finish generating one draft note to be around 2 hours. A total of 1633 actions (ADD/UPDATE/REMOVE) were recorded with 152 (9%) UPDATEs and 0 REMOVE. The lack of REMOVE action taken by the scribes is expected as in most cases removing medical content from a note would be accompanied by the addition of new information, which would then be captured by UPDATE actions. However, we decided to keep REMOVE as a supported action in the task definition and guidelines to account for unexpected future cases. The average number of evidences marked for each note item is 3.4 lines, which aligns very well with the objective of incrementally adding new information instead of summarizing large body of text every time; since the guidelines do not specify any constraints on the number of evidences per note item, this shows that our scribes understood well the incremental nature of the task and adapted their decision making process accordingly.

4.2. Inter-rater agreement

Inter-rater agreement is a focus of our analysis on the annotated data. We treat marked evidences as a 1/0 labels on all lines in a conversation, and utilize Cohen's κ as the metric for inter-rater agreement; since κ is not defined for comparing texts, we leverage ROUGE scores (Lin, 2004) to measure the agreement, or similarity, between the note texts. Specifically, the reported ROUGE-1/2/L F1 scores (between 0 and 1) effectively measure the percentage of common unigram/bigram/longest substring between two texts.

We report text similarity (Table 4) and κ on marked evidences (Table 5) at three different levels: full note, category, and item. At full note level, the similarity scores (ROUGE or κ) between two draft notes are calculated and averaged across all pairwise comparisons among scribes; updated

note items are ignored and repeated evidences are counted only once. At category level, each draft note is first divided by sections into three categories: Subjective, Objective, Assessment/Plan, and then the same metrics are calculated and averaged in the same way between the same category of two notes. Item-level similarity scores are not obtained as straightforward: there exists no natural mapping between note items created by different scribes, and different scribes can create different numbers of note items or use different sets of sections; therefore, we approximate the similarity by pairing each item in one note with the one item in a second note that is (i) in the same category and (ii) has the most similar set of marked evidences by κ[3]. The scores are then calculated and averaged across all paired items. This approximation means the item-level ROUGE scores in Table 4 is an optimistic estimation of how similar two notes are across individual sentences; while κ in Table 5 sets an upper bound on the agreement in the marked evidences between "similar" items in two notes. Both tables also report separate scores obtained from the training phase and the pilot phase. The key findings are discussed as follows:

	Training phase	Pilot phase
Full note	0.410/0.147/0.221	0.430/0.158/0.264
Category level		
Subjective	0.419/0.151/0.238	0.378/0.144/0.257
Objective	0.217/0.095/0.154	0.354/0.140/0.293
A/P	0.301/0.113/0.206	0.311/0.120/0.226
Item level		
Subjective	0.320/0.137/0.279	0.360/0.148/0.327
Objective	0.210/0.070/0.189	0.313/0.127/0.282
A/P	0.231/0.075/0.175	0.315/0.121/0.269

Table 4: Draft note similarity by ROUGE scores measured at different granularity. Each cell displays the ROUGE-1/2/L F1 scores, averaged across all pairwise comparison among four scribes and all conversations. Higher values mean higher similarity.

Training is necessary and effective. Almost all metrics show improvement from the training phase to the pilot phase, with the agreement in the marked evidences showing much greater improvement than text similarity. This generally aligns with our training regimen that focuses more on correcting and clarifying the procedure of the task than on restricting the writing style or word choice.

Text similarity improves at note level. This aligns with our desired outcome: while we do not place explicit constraints on the language of the text, we do expect the information to be complete in the final notes. A higher ROUGE score at note level compared to item level is therefore an indication of a better agreement among scribes on the medical content annotated in the full note, even though the allocation of the content may be less consistent.

Consistency across notes could be underestimated by the reported values. We calculated the typical ROUGE between two randomly chosen notes. The ROUGE-1/2/L F1 scores are around $0.24/0.06/0.16$, respectively. Therefore, the reported values in Table 4 are up to 100% higher than the level of similarity between random notes created on conversations of the same specialty, indicating that the semantic similarity across notes from different scribes is probably much higher than the face value. Additionally, we calculated the Pearson correlation coefficient between κ and ROUGE at item level and obtained $0.998/0.967/0.998$ for all paired note items in Subjective/Objective/Assessment&Plan, respectively. This means that when two scribes mark very similar evidences, the text they use to summarize the content will also be highly similar.

Inter-rater agreement varies significantly across categories. Text similarity is lower in Objective and Assessment&Plan categories compared to Subjective category. These two categories also show much bigger improvement via training. We believe this is a reflection of the intrinsic variation in how humans filter information into different sections of a SOAP note. Subjective sections like HPI require coherent and narrative text and experienced scribes are usually extensively trained in the writing style; whereas Objective sections such as PE contain more itemized information and different physicians may have different preferences on the template and the scope of information to be included in a SOAP note, thereby leading to a variation in the decision process of different scribes. By contrast, the κ is generally higher in the Objective category, which is reasonable considering that evidences selected for itemized information (such as items in the PE section) tend to be more direct or verbatim and therefore easier to be agreed upon between scribes.

It is worth pointing out that a target level of agreement is yet to be established due to the complexity and novelty of the ING task. We therefore believe that the results presented in Table 4 and 5 establish a reference baseline for future annotation work on the topic of generating SOAP notes.

The feasibility of the EMU task is supported by results in Figure 3. Figure 3(top) shows the detailed

[3]If all κs are zero, i.e. no note item in the second note shares any evidences with the first item, then we map it to the item that yields the highest ROUGE-1 F1 score.

	Training phase	Pilot phase
Full note	0.220	0.392
Category level		
Subjective	0.348	0.398
Objective	0.184	0.499
A/P	0.265	0.238
Item level (upper bound)		
Subjective	0.419	0.426
Objective	0.216	0.427
A/P	0.631	0.683

Table 5: Inter-rater agreement on the marked evidences at different granularity. Each cell displays Cohen's κ averaged across all pairwise comparison among four scribes and all conversations. The item level κ is approximated as an upper bound on inter-rater agreement, see Section 4.2 for detailed explanation.

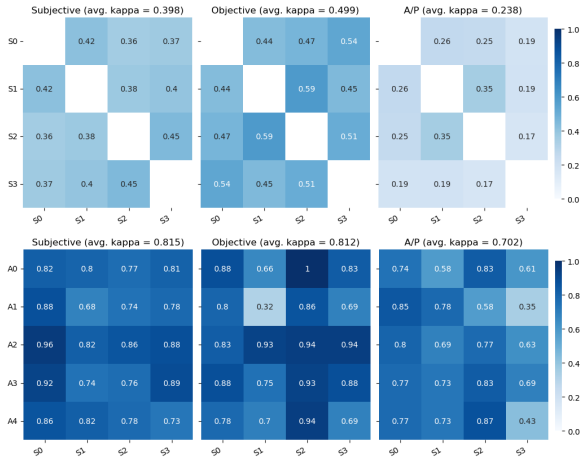

Figure 3: Inter-rater agreement (Cohen's κ) on marked evidences at category level: (top) between scribes; (bottom) between EMU annotators and scribes. [*S0...S3*] index the four scribes and [*A0...A4*] represent the five EMU annotators.

breakdown of the inter-rater agreement between four scribes on the marked evidences at category level (4th through 7th rows in Table 5), while Figure 3(bottom) shows the same metric measured between evidences marked by EMU annotators, or the linguists, and the ground-truth evidences marked by the scribes. Note that only the last marked evidence in each note item is shown to the annotators during EMU. They need to recover all other evidences in the portion of the transcript up to the last marked evidence. It is interesting that the agreement is much higher in the EMU step than between scribes in the Generation step. Admittedly, the EMU step rather resembles the relatively easy information retrieval than text generation; the reported high κ nonetheless strongly suggests the EMU step is achievable to high accuracy by non-medical experts and thus beneficial for scaling up the ING task.

4.3. Case study on the Rewriting step

Given the relatively simple objective of the Rewriting step and the limited data we collected during the pilot study, we choose to conduct case studies on the changes scribes made during the rewriting. In Figure 4, we visualize one such study as a "similarity bipartite map" between a draft note and its rewritten version with edges that connect sentences of high similarity. The labeled numbers are the average of ROUGE-1/2/L F1 scores between the connected sentences, which we use as a proxy for similarity measure; the rewritten note is treated as the reference so that every sentence within the note gets connected from only one sentence of the highest average ROUGE score in the draft note.

We can see in the example the necessity of the rules we outlined for Rewriting in Section 2.2. The reordering of sentences and reformatting (e.g., changing "MEDS" into "Medications" as section header) are in particular frequent as indicated by the number of crossing edges in the map. Distributing and merging information[4] even across sections are frequently present during rewriting, e.g. the first sentence in Imaging & Labs section in the draft note gets mapped to two sentences in the rewritten note. The specific editing history during rewriting could provide additional structured data for modeling purposes, which we consider as a future direction and discuss briefly in Section 5.

5. Discussion

We would like to reiterate the valuable experience we accumulated through the designing of the ING framework and the pilot study. Almost all aspects of the framework (GUI, guidelines, etc.) are the results of an iterative and direct collaboration with the intended annotators for the task: the medical scribes. We believe the level of involvement of actual annotators contributed significantly to making such a complicated task feasible and practical. In the meantime, we also learned several lessons in regards to the challenges of ING as an annotation task:

First, medical expertise and scribing experience do not translate as expertise in the ING task. The proposed framework shifts from the normal workflow of a medical scribe and we have observed that significant deviance and variation from our intended annotation can occur without training. But as is shown in Section 4, properly designed training regimen through collaborative annotation, evaluation against reference, and feedback sessions can

[4]We use "information" instead of "sentences" because this is inferred from the proxy similarity scores. We do not know if scribes actually conducted splitting, merging, or copy-paste of text during rewriting.

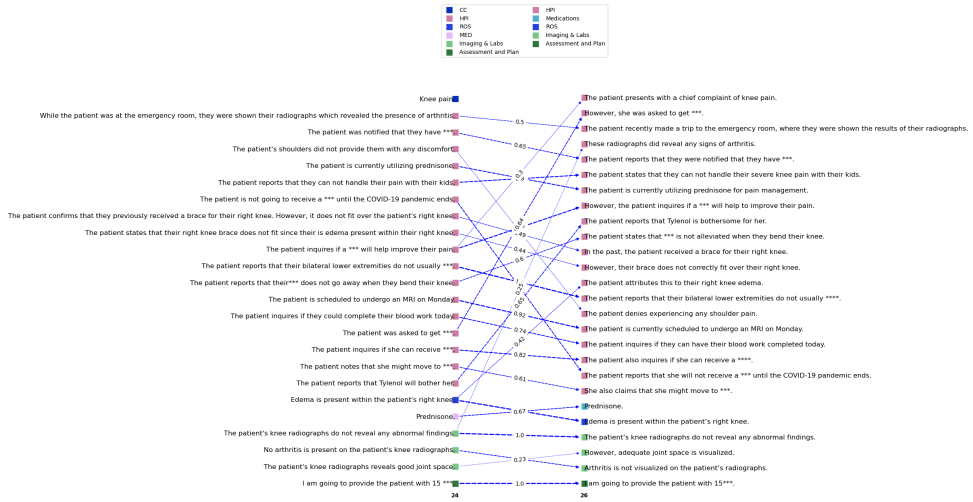

Figure 4: Similarity bipartite map on a rewriting job. The sentences in the final draft note are listed in the order of their creation on the left, while on the right lists all ordered sentences in the rewritten SOAP note. Section headers in both versions are subsumed into the colors of individual nodes. The numbers on each edge/link in the map shows the average of ROUGE-1/2/L F1 scores calculated between the connected sentences.

prepare medical scribes for the task in reasonable time and effort.

Second, variation is inevitable in the output. Difference in the level of medical expertise and prior scribing experience with different healthcare providers introduce unavoidable variation in the summarization styles between different scribes. Even outside the medical domain, the thought process of a human summarizing a conversation is a highly stochastic process. This means the output data from the ING task will be as diverse as the number of annotators involved. However, we consider this diversity as a value more than a challenge for training robust or interpretable deep neural networks (potentially through reinforcement learning).

Third, quality control is difficult. We have observed some scribes making factual mistakes during the training phase and had to shift the focus of some of the feedback sessions to instructing on how to summarize information for a SOAP note. Though not intended to be part of the ING task training, it does reflect the inherent difficulty of SOAP note generation even for experienced medical scribes. A more relevant implication is that quality control for the ING task will involve both the final note (draft or rewritten) and the intermediate "thought process", and we have yet to reach a standard quantitative measure of quality for the latter. However, the results in Section 4.2 do demonstrate that a reasonable level of consistency can be attained even between the incremental processes of different scribes given a proper understanding of the task.

Future direction Scaling up the ING task on SOAP note generation is a natural next step for us. As of writing of this paper, we have already started a large scale ING task on 1000 DoPaCos. We are also constantly updating the ING GUI to embed more task requirements and quality control as features (e.g. adding NLP models to check and flag problematic note items during the annotation), thereby further simplifying guidelines and training. The Rewriting step also offers promising expansion of the ING task, as we show through the case study in Section 4.3: currently the only output from the Rewriting step is the rewritten note; however, with a properly designed GUI, we can incorporate more data in the annotation that captures the editing history of a scribe's rewriting process, including the reordering, splitting, and merging of note items.

6. Conclusion

We propose the novel framework of Incremental Note Generation as a first attempt to capture the process of summarization as an annotation task. By decomposing the process of conversation summarization into incremental generation steps that aim to produce an ordered series of accumulative draft notes, and a following rewriting step to polish the final draft note into a reference note, we show that the decision making process during the creation of the final reference note can be annotated as tabular data with a properly designed column structure. The pilot study on the SOAP note generation from doctor-patient conversations demonstrated the feasibility of the task and a reasonable level of consistency in all aspects of the annotated

data. Although the pilot study was conducted in the medical domain, we believe the ING task serves as a valuable annotation framework for dialogue or long text summarization in general, and the rich structure of the resulting data balances or even outweighs the added complexity compared to a traditional annotation workflow for summarization.

7. Ethical Statement

The corpus used in this study includes transcripts of encounters between doctors and patients. All transcripts have gone through a thorough anonymization and de-sensitization process to remove all PHI (personal health information) contents. SOAP notes are created solely based on the de-PHIed version of the transcripts. All procedures are HIPAA compliant and risk of exposing PHI information to the public is minimal. Scribes and annotators employed in the annotation project are well compensated by the hour.

8. Acknowledgements

We thank Melinda Childs and Dr. Adam Rothschild for their many constructive discussions and contribution to the detailed guidelines. We also deeply appreciate the time and effort spent by all our expert (scribes) and non-expert annotators.

9. Bibliographical References

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Jean Carletta. 2006. Announcing the ami meeting corpus. *The ELRA Newsletter*, 11(1):3–5.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Shen Gao, Xin Cheng, Mingzhe Li, Xiuying Chen, Jinpeng Li, Dongyan Zhao, and Rui Yan. 2023. Dialogue summarization with static-dynamic structure fusion graph. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13858–13873, Toronto, Canada. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Huan He, Sunyang Fu, Lifwei Wang, Sijia Liu, Andrew Wen, and Hongfang Liu. 2022. Medtator: a serverless annotation tool for corpus development. *Bioinformatics*, 38(6):1776–1778.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 385–394, Seattle, United States. Association for Computational Linguistics.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Amanda Perry, Ashley Robson Domin, Chris Co, Gang Li, Hagen Soltau, Izhak Shafran, Justin Stuart Paul, Lauren Keyes, Laurent El Shafey, Linh Tran, Mark David Knichel, Mingqiu Wang, Nan Du, Rayman Huang, and Yu hui Chen. 2020. The medical scribe: Corpus development and model performance analyses. In *Proc. Language Resources and Evaluation, 2020*.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2022. SOAP notes.

Ashwyn Sharma, David Feldman, and Aneesh Jain. 2023. Team cadence at MEDIQA-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 228–235, Toronto, Canada. Association for Computational Linguistics.

Jing Su, Longxiang Zhang, Hamid Reza Hassanzadeh, and Thomas Schaaf. 2022. Extract and abstract with bart for clinical notes from doctor-patient conversations. *Proc. Interspeech 2022*, pages 2488–2492.

Ye Wang, Xiaojun Wan, and Zhiping Cai. 2022. Guiding abstractive dialogue summarization with content planning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3408–3413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Wen-wai Yim, Meliha Yetisgen-Yildiz, Jenny Huang, and Micah Grossman. 2020. Alignment annotation for clinic visit dialogue to clinical note sentence language generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 413–421.

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

A. Detailed ING Guidelines

This section elaborates the guidelines in the original writing we share with the scribes and annotators:

A.1. Project Overview

The ING task is meant to collect medically relevant evidences from a conversational transcript and transform them into note items in an electronic health record SOAP note. This is done by selecting relevant utterance lines and writing up a note item in the corresponding SOAP note category. The task is also meant to capture the point in the transcript that gave justification for a change (information adding or modifying) in the note and when this change occurred. To accomplish this, the information is collected in a linear fashion making updates to previously made notes as necessary as the annotator progresses through the transcript.

A.2. Term Definition

Utterance: a collection of speech from an individual speaker during an encounter, separated by at least 0.3 seconds of silence or by punctuation.

Evidence: each individual piece of information within an utterance used to create a note item (Please note: An individual utterance may contain evidences for multiple note items).

Note Item: each individual piece of information generated by an annotator from the evidences within the utterances of the encounter.

A.3. Format

- Note items should be as concise as possible while still capturing all the relevant dependencies (i.e. if a patient has stopped exercising regimen due to shoulder pain, the note should not have two separate items of “patient stopped exercising” and “patient presents with shoulder pain”. Instead, because the exercise being discontinued because of the shoulder pain, both pieces of information will be included in a single note item).

- Abbreviations used in the conversation transcript will be expanded in note items, unless they are a part of the Approved Abbreviations List found below.

- The note items being generated are pieces of a note, so they can have a bullet point style that will be processed for readability in a future step. The goal is to include all the necessary information, so that a comprehensive note can be compiled from the items generated in this task.

- Avoid adding in inferred information. For example, it is often added that a patient is agreeable to a plan at the end of the note. However, unless there is interaction between the patient and provider that states that this is true, do not assume. IE: provider asks “how does that sound?” and patient responds “good” then this would be appropriate. Without this interaction, then there is no need for a line stating “patient is agreeable to plan”.
- Utilize the SOAP note sections as you would in a typical live scribing scenario.

A.4. Process

- As the task is meant to be completed linearly, annotators will create note items when the minimal requirements necessary to add/update/delete a note item are present (e.g. sufficient evidence to support any note item being added or modified including dependencies if present). This may include simple yeses and nos to questions being asked. If a note item is complex and spread over multiple groups of lines throughout the file, it can be broken up into separate add and update items as applicable.
- The update action is meant to allow annotators to take in historical context, building on the information they have encountered prior in the transcript. This is meant to simulate the workflow of a synchronous scribe, who would only be able to listen to the conversation in linear order and would thus have to modify the note as new information becomes available.
- As the focus of the task is on note creation not correlation, repeated information that has already been documented will not be captured unless the information changes some part of the previously selected items. Note: Dictation by the provider will supersede prior note items and should update the corresponding note items verbatim where applicable.
- ING vs ING+EMU
  - ING: During the note generation portion of the task, the annotator will select all evidences that are necessary to support the note item being added/updated/removed and does so for each subsequent action.
  - ING+EMU: During the note generation portion of the task, the annotator will select only the final line of evidence (lower boundary) of the evidences necessary to support the note item being added/updated/removed and does so for each subsequent action. In the EMU portion of the task the annotator will go through the transcript and mark all evidences for each generated note item/action within the given boundaries.

Example	Explanation
*Patient states having experienced depression symptoms for ****	We are unsure of just how long the depression symptoms have been going on as the audio was garbled at this moment.
*Patient's last pap smear was *** with normal findings*	The pap smear's exact date was given per provider preference and then removed as this would be PHI.

Table 6: Example of unknown information annotation.

A.5. Unknowns

For any unknown information that may be missing due to protected health information or by garbled audio, utilize *** in its place. See Table A.5 for examples.

A.6. SOAP Basics

This part is meant to provide an overview of the larger categories of the SOAP note. Not all sections are included in these definitions as we rely on scribe experience to determine the appropriate sections for a given note item.

Subjective includes but not limited to chief complaint, history of present illness, and review of systems.

CC (Chief Complaint)

- The chief complaint will be whatever the main reason for the visit is or the initial reason. This may not be explicit but will be most notable throughout the visit. Sometimes it can be as simple as the provider telling the scribe "follow up of *** y/o/disease" or "annual wellness exam/physical" or more detailed depending on the issues for the patient.
- Does not need to be a full sentence.

HPI (History of Present Illness)

- Try to tell a story (as best as possible with limited information).
- Use phrases like “patient states, endorses, denies, reports, complains of, etc.” and avoid

phrases that can be found to be false. Consider "patient had right shoulder arthroplasty" is different from "patient reports having right shoulder arthroplasty" the second is a better choice for subjective information.

- Avoid using specific gendered pronouns unless the provider uses them to scribe or to refer to the patient and use "patient" and third person singular form for verbs until pronouns are confirmed by the provider. (Reminder: Just because someone is getting a PSA does not mean they use he/him pronouns).
- Separate different diagnoses from each other.
- Not all topics discussed will be part of the HPI. Context and specialty can inform the decision to include or omit certain details IE: an issue discussed in Family medicine is more often relevant than if the same issue was discussed in a less broad specialty.

ROS (Review of systems)

- Usually, a list of questions that can be related to the illness(es) the provider is trying to diagnose or just general information on the patient.

Objective includes but not limited to physical exam, labs, and imaging

PE (Physical Exam)

- When working in a specialty like orthopedics, dermatology, plastics, likely not all exams will be needed and just the ones the provider mentions.
- For physical exams, nothing will be added unless it is explicitly stated by the provider. Below is a standardized “normal” exam list. If a provider mentioned that an area is normal, please add the line below that correlates to that area, adding and deleting as needed, otherwise only add in information that is dictated.
- Standardized Exams
  - General: Well-appearing, NAD
  - Heart: RRR, no bruits, clicks, or rubs, no murmurs present
  - Lungs: Clear to auscultation bilaterally
  - Abd: no guarding or rebound
  - Extremities: no pitting edema
  - Skin: no rashes or erythema
  - Psych: Ox4, no SI or HI

Labs, Imaging, etc.

- If any labs or imaging are present, include the name, location, and date listed in the note item. For any missing information put ***.
- List whatever information the provider states to the patient or scribe about their x-ray.

Assessment and Plan includes but not limited to A/P.

A/P (Assessment and Plan)

- For this section, take the diagnosis and write it as "diagnosis: plan to follow". If the patient is diagnosed with diabetes, then told to follow up in 3 to 6 months for an A1c after working on their diet and exercise, then write it like this and update the information as you find more throughout the conversation.
- Dictation outweighs all other information. Especially in the A/P. Dictation is typically an addition to HPI.
- Be as detailed as possible.
- Use only the abbreviations in the Approved Abbreviations List below.

A.7. Approved Abbreviations List

Table 7 - 9 list all approved abbreviations.

Abbreviation	Definition
DMII	Diabetes Mellitus Type 2
SOB	Shortness of Breath
HTN	Hypertension
MI	Myocardial Infarction
STEMI	ST-Elevated MI (ST segment of EKG is spiked, indicating heart attack)
PE	Pulmonary Embolism
CVA	Stroke
COPD	Chronic Obstructive Pulmonary Disease
DVT	Deep Vein Thrombosis
FX	Fracture
URI	Upper Respiratory Infection
UTI	Urinary Tract Infection
CHF	Congestive Heart Failure
NIDDM/IDDM	Non-Insulin Dependent Diabetes/Insulin Dependent Diabetes
SBO	Small Bowel Obstruction
HLD	Hyperlipidemia

Table 7: Approved Abbreviations - Conditions

Abbreviation	Definition
BUN	Blood Urea Nitrogen
CBC	Complete Blood Count
TSH	Thyroid Stimulating Hormone
H&H	Hemoglobin and Hematocrit
CMP	Comprehensive Metabolic Panel
A1c	Hemoglobin A1c
UA	Urinary Analysis
BMP	Basic Metabolic Panel
INR	International Normalized Ratio
ABG	Arterial Blood Gas
LDL	Low-Density Lipoprotein
HDL	High-Density Lipoprotein

Table 8: Approved Abbreviations - Labs

Abbreviation	Definition
QD	Everyday
PRN	As Needed
PO	By Mouth
OTC	Over the Counter

Table 9: Approved Abbreviations - Others

B. UI for Evidence Mark-Up

The UI (as of writing of this paper) for Evidence Mark-up step (Section 2.3) is shown in Figure 5 (navigation panel and page title omitted). The main components between the EMU UI and the ING UI (Fig. 2) are largely identical, except for two added components/features (indexed as in the figure): (4) previous note item preview. This is a non-editable textbox showing any annotation items that are updated by a later item, i.e., this box shows the "before" state of an UPDATE action; (5) most of the transcript area is greyed out, this is to help the annotators quickly locate "focus" region in the transcript for potential evidences; these regions are calculated based on the bounds between the currently selected annotation and the previous annotation; in case of an updated item, the bounds are determined between the currently selected annotation and the item before the update.

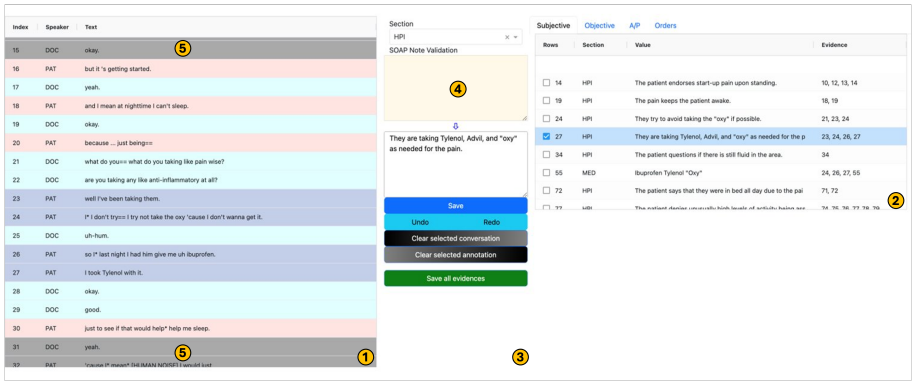

Figure 5: GUI for EMU task. Indexed components are: (1) transcript table; (2) note item table; (3) annotation tools; (4) previous note item preview; (5) greyed out transcript region.