Learning functional sections in medical conversations: iterative pseudo-labeling and human-in-the-loop approach

Mengqian Wang* MENGQIAN@ALUMNI.UNC.EDU
University of North Carolina at Chapel Hill
Ilya Valmianski† ILYA@AUXHEALTH.IO
AuxHealth
Xavier Amatriain XAVIER@CURAI.COM
Anitha Kannan ANITHA@CURAI.COM
Curai Health

Abstract

Medical conversations between patients and medical professionals have implicit functional sections, such as "history taking", "summarization", "education", and "care plan." In this work, we are interested in learning to automatically extract these sections. A direct approach would require collecting large amounts of *expert annotations* for this task, which is inherently costly due to the contextual inter-and-intra variability between these sections. This paper presents an approach that tackles the problem of learning to classify medical dialogue into functional sections without requiring a large number of annotations. Our approach combines pseudo-labeling and human-in-the-loop. First, we bootstrap using weak supervision with pseudo-labeling to generate dialogue turn-level pseudo-labels and train a transformer-based model, which is then applied to individual sentences to create noisy sentence-level labels. Second, we iteratively refine sentence-level labels using a cluster-based human-in-the-loop approach. Each iteration requires only a *few dozen* annotator decisions. We evaluate the results on an expert-annotated dataset of 100 dialogues and find that while our models start with 69.5% accuracy, we can iteratively improve it to 82.5%. Code used to perform all experiments described in this paper can be found here: https://github.com/curai/curai-research/functional-sections.

Keywords: Medical NLP, Medical Dialogue, Medical Sections, Pseudo-labeling, Human-in-the-loop

1. Introduction

Recent growth in telemedicine has led to a dramatic expansion in text-based chat communications between patients and medical professionals (Bestsennyy et al., 2021). This creates new opportunities for improving medical professional workflows through the introduction of natural language understanding (NLU) systems for providing real-time decision support and automating electronic health record (EHR) charting (Dreisbach et al., 2019; Joshi et al., 2020; Valmianski et al., 2021). Auto-charting, in particular, benefits significantly from proper contextualization of the dialogue (Khosla et al., 2020; Krishna et al., 2021). For example, the History of Present Illness (HPI) section of the progress note can

* Work done during an internship at Curai Health
† Work done during working as a full-time Machine Learning Researcher at Curai Health

be derived from the history-taking discussion in the dialogue, while the Care Plan section can be derived from the care plan discussion.

Figure 1: A de-identified patient-medical professional dialog color-coded with the predictions from our approach (best seen in color). Notice how different multiple labels may be present in the same turn of the conversation.

Similar to medical SOAP note (Podder et al., 2021) sections, we consider labeling each medical professional written sentence into the following functional sections:

1. *History taking*: questions about the patient's current illness including symptoms, prior medical history, and medications they may be taking.
2. *Summarization*: medical professional's confirmation of relevant patient symptomatology.
3. *Education*: education of the patient about their medical issues.
4. *Care plan*: suggestion on the course of action or treatments.
5. *Other*: non-medical text.

Figure 1 shows an abridged dialogue that is color-coded with appropriate section labels, as predicted by the model introduced in this paper. We make two observations. First, within a single turn of the dialogue, multiple functional sections (or classes) can co-occur *e.g.* educating the patient that purple feet are a diabetes-related symptom while taking history on whether the patient has been previously diagnosed with diabetes. Second, dialogue sentences belonging to different functional sections may have high lexical overlap, *e.g.* purple feet being discussed in the context of history taking, summarization, and education.

We formulate the problem of inferring conversation functional sections as sentence-level classification: *Given a medical professional-patient dialogue, how can we assign every sentence in every turn of the dialogue to the correct functional section?* Further, *How can we learn such a model when we can get only small amounts of human-generated labels?*

We tackle both these questions by leveraging two key insights:

1. *Dialogue turns carry more information than individual sentences and are thus easier to learn with weak supervision.* We use this insight to build a noisy turn-level labels dataset and train a language model to classify turn-level labels. We then apply the turn level model to label individual sentences within the turn, creating noisy sentence-level labels.

2. *Text embeddings learned from noisy labeled data are relevant even when the classifications are not reliable.* We use this insight to propose an iterative human-in-the-loop cluster-based pseudo-labeling strategy. Our proposed clustering strategy introduces

variability in samples across iterations by enabling intermixing high-confidence predictions with low-confidence ones and choosing only class-specific 'pure' clusters through a simple human-in-the-loop evaluation.

We evaluate the results on an expert-annotated dataset of 100 dialogues and find that although the initial pseudo-labels have an accuracy of 69.5%, our iterative refinement approach can boost accuracy to 82.5%. We also find that the latent space representations of each class become both more tightly clustered and more separable between different classes, which may imply higher generalizability (Li et al., 2020).

2. Generalizable Insights about Machine Learning in the Context of Healthcare

Healthcare datasets often suffer from insufficient labeling. To effectively classify medical data, categories must be functional, and the labeling process typically demands costly subject matter experts (SMEs). Our strategy involves fine-tuning pretrained models by using a minimal amount of SME-labeled data in an iterative human-in-the-loop fashion. During each iteration, we embed and cluster raw medical conversation data, discarding low-purity clusters in the following training iteration. This method uses minimal human input and has led to a substantial model performance improvement. The approach outlined in this paper is applicable to any type of categorizable textual data, offering value by lowering data labeling expenses.

3. Related Work

Semantic structure understanding: The importance of identifying and assigning labels to functionally coherent units is well-understood. As an example, in legal document understanding, Saravanan et al. (2008); Malik et al. (2021) show that it's easier for downstream tasks if documents are segmented into coherent units such as facts, arguments, statutes, *etc.* In conversational dialogues, the problem of utterance-level intent classification to detect discourse boundaries is well studied (Liu et al., 2017; Raheja and Tetreault, 2019; Qu et al., 2019; Joty et al., 2014; Takanobu et al., 2018). These intents are broad (*e.g.*"original question" and "repeat question" Qu et al. (2019)) and identified at turn-level.

We are interested in classifying dialogue turns and also each sentence within a turn into functional sections (history taking, summary, education, care plan, other) that can loosely serve as intents. These sections interleave (*e.g.* history taking and education) within a single dialogue turn making the task challenging. Previous works assume access to manually labeled data. In this paper, we bootstrap data using a weak pseudo-labeler and then iteratively refine it with training text-classification models, clustering their embeddings, and relabeling entire clusters using a human-in-the-loop.

Active learning: This approach focuses on starting with a small labeled dataset and iteratively retraining models with an updated labeled dataset (see references in survey papers Settles (2009) and Ren et al. (2020)). Each update to the labeled training set involves getting manual labels for a small (often only one) number of most informative examples - examples of which the model at the previous iteration is most uncertain.

In contrast, our human-in-the-loop approach aims to change labels in a much larger number of examples in each turn. We cluster the embeddings of the examples based on the current model, get cluster-level annotation from the human annotators, and impute that label to all the examples of the cluster. Of related is the work of Mottaghi et al. (2020) that uses clustering within the active learning framework, but the technique was used to only identify previously unseen classes and to obtain a small number of informative examples within each cluster to increase coverage.

Pseudo-labeling: Another approach to impute labels to a large number of unlabeled examples is to (Mindermann et al., 2021; Du et al., 2020; Chen et al., 2020) use a trained model's prediction to self-label (self-training or pseudo-labeling hyun Lee (2013)).

Du et al. (2020) shows that self-training with pseudo-labeling can improve performance on text classification benchmarks without the need for in-domain unlabeled data. While being general and domain-agnostic, pseudo-labeling approaches can under-perform if the generated labels are noisy (e.g., high variance model in the previous iteration of training) and hence adversely affect performance (c.f. (Oliver et al., 2018; Rizve et al., 2021; Nair et al., 2021) and references therein). In this paper, we combine pseudo-labeling followed by independent clustering of the pseudo-labeled-class specific data points. Human experts then annotate samples from each cluster to either relabel the entire cluster or remove it from the next training iteration (because it contains sentences from multiple functional sections).

4. Approach

In this section, we present a general description of our approach. We describe the specifics of applying this approach to medical conversations in § 5.

Figure 2 presents a schematic overview of our approach. It consists of two parts. First, in turn-to-sentence label bootstrapping, we pseudo-label turn-level labels [1] which we use to train a text classification model. We then apply this model to sentences to create noisy sentence-level labels (§ 4.1). Second, we iterate on sentence-level labels by training a text classification model, clustering the sentence-level embeddings, and then using a human-in-the-loop to classify the sentences of each cluster (§ 4.2). The notation used in this paper is described in Table 1.

4.1. Turn-to-sentence label bootstrapping

In this step, we train a turn-based model (Figure 2b) to serve as the noisy pseudo-labeling for sentence-level labeling. We use a set of weak labelers to generate a turn-level multilabel dataset for this task (Figure 2a) of the form: $L_i^{\text{turn}} = \cup_j L_{ij}$.

L^{turn} are used to train a turn-level multilabel model M_{turn} (Figure 2b). M_{turn} is then used to generate sentence-level labels by applying directly on sentences instead of entire turns, $L_{ij}^0 \leftarrow M_{\text{turn}}(S_{ij}, D)$ (Figure 2c). Note that in § 5.2 we discuss how, in our application, we still find it useful to apply (simple) rules on top of trained model output.

1. We use labels and functional sections interchangeably based on the context

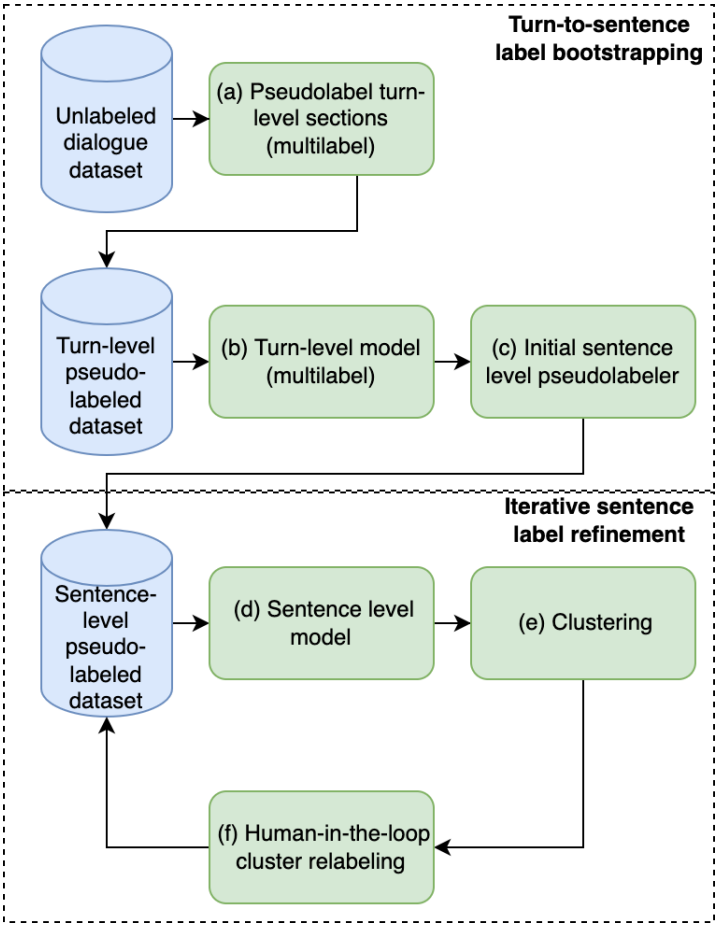

Figure 2: Schematic of the approach. The output of this approach is both the labeled dataset (after step f), and a classification model (step d).

Symbol	Description
D	Dialogue
T_i	i'th turn of the dialogue
S_{ij}	j'th sentence in the i'th turn of the dialogue
$\mathcal{L}$	Universe of labels
L_i^{turn}	Turn level functional section label of the i'th turn
L_{ij}^k	k'th iteration sentence level functional section label of i'th turn and j'th sentence
M_{turn}	Text classification model trained on turn level functional section labels
M_{sent}^k	k'th iteration text classification model trained on sentence level functional section labels
$\hat{L}_{ij}^k$	Estimated labels for S_{ij} produced by M_{sent}^k
E_{ij}^k	Embedding of the S_{ij} by M_{sent}^k
Clst	Clustering algorithm applied independently to embeddings E for each functional section in $\mathcal{L}$
C_{ij}^k	Cluster assigned to S_{ij} by Clst using embeddings and labels produced by M_{sent}^k.
H	Human annotator that reviews a set of sentences and assigns the set one of the labels in $\mathcal{L}$ or marks them as "Mixed"
L_n^{clst}	A label applied to all sentences of cluster n (*e.g.* $C_{ij}^k = n$)

Table 1: Notation used in this paper.

4.2. Iterative sentence label refinement

The iterative refinement starts with training a sentence-level text classification model (Figure 2d), which is then used to produce both estimated labels and embedding $(\hat{L}_{ij}^k, E_{ij}^k)$. These embeddings are clustered independently for each functional section label (Figure 2e), and then each cluster, as a whole, is relabeled by a human annotator (Figure 2f). Examples from clusters that contain sentences belonging to different functional sections (as judged by a human annotator and marked as "Mixed") are not used in the next iteration of retraining. See algorithm 1 for details.

Input : Dialogue dataset $\mathcal{D}$
Current iter. model M_{sent}^k
Current iter. pseudo-labels L_{ij}^k
Clustering algorithm Clst
Cluster annotator H
Output: $\{L_{ij}^{k+1}\}$

$\{(\hat{L}_{ij}^k, E_{ij}^k)\} \leftarrow$
$\{\forall D \in \mathcal{D}, \forall S_{ij} \in D, M_{\text{sent}}^k(S_{ij}, D)\}$

$\{C_{ij}^k\} \leftarrow \text{Clst}(\{(\hat{L}_{ij}^k, E_{ij}^k)\})$

$\{S_{ij}\}_n \leftarrow \text{Sample}(\{S_{ij} : C_{ij}^k = n\})$

$L_n^{\text{clst}} = H(\{S_{ij}\}_n)$

$L_{ij}^{k+1} \leftarrow L_n^{clst} : n = C_{ij}^k$

return $\{L_{ij}^{k+1} : L_{ij}^{k+1} \neq \text{Mixed}\}$

Algorithm 1: Pseudocode for iterative cluster refinement of sentence level models. Sample function draws a small number of examples (we found 10 examples to be the smallest but the most efficient number for our dataset, which could differ in other datasets) from a set. "Mixed" represents that the set of sentences has sentences that pertain to several different functional sections (greater or equal to two out of ten in our dataset).

5. Experimental details

5.1. Dataset

We use a dataset with 60,000 medical professional-patient encounters containing over 900,000 dialogue turns and 3,000,000 sentences collected on a telehealth virtual primary care platform. To construct a test set, we randomly sampled 100 encounters (not used for training or validation) for which we procured human labels for all medical professional written sentences (3,102 sentences). In the human-labeled dataset, the distribution of sections on

the sentence and turn levels are respectively: summarization: 3.6%, 2.6%; history taking: 26.5%, 31.7%; education: 5.3%, 8.4%; care plan: 4.1%, 7.9%; other: 60.3%, 49.3%. We do **not** have any additional labels for these encounters.

5.2. Turn-to-sentence label bootstrapping

As described in § 4.1, we first generate a dataset using *ad hoc* methods to train a turn level multilabel classification model. We then use this model to pseudo-label individual sentences.

Unsupervised clustering and human annotation of clusters. We embed dialogue turns into fixed-sized representations by mean-pooling the final layer of the off-the-shelf DeCLUTR[2] (Giorgi et al., 2021) sentence encoder. Following Allaoui et al. (2020), we project the 768D original embedding space to 250D via PCA and then project via UMAP (McInnes et al., 2018) to 50D. We then cluster these 50D representations using the k-means++ algorithm (Arthur and Vassilvitskii, 2007) and determine the number of clusters using the elbow method (Thorndike, 1953) (in our dataset, this number was 10). Human annotators manually label the resulting clusters by examining a small number ($\sim$ 10) of sentences in each cluster Figure B.1.

Human annotation of a cluster-derived set of examples. Because the unsupervised clustering did not produce good clusters containing only education or care plan turns, we procured human labels for 5000 turns from a mixed cluster containing education and care plan turns.

String-based rules. We identify turns with summarization sentences by string matching one of ['summar', 'sum up'].

Turn-level model to generate sentence pseudo-labels. We construct the dataset for the turn-level model by assigning the same label as the cluster after removing all mixed clusters. We then train M_{turn}, a multi-label classifier on top of DeCLUTR using this turn-level labeled set. The classification head consists of a single feed-forward layer with sigmoidal activation for each label.

To create the initial sentence level labels, we apply the turn-level model on each sentence and assign labels according to algorithm 2 in the Appendix.

5.3. Iterative sentence label refinement

Sentence-level model. The input to this model is the dialogue turn that contains the target sentence. We mark the target sentence with tokens ⟨START⟩ and ⟨END⟩. The model itself consists of a transformer language model DeCLUTR sentence encoder, with a classification head consisting of a single feed-forward layer with a softmax activation.

2. We also tried BioBERT (Lee et al., 2019), Mirror-BERT (Liu et al., 2021), and Sentence BERT (Reimers and Gurevych, 2019), but found that DeCLUTR produces representations that cluster with high label-purity

	Round transition		
	1→2	**2→3**	**3→4**
History taking	-	1/10→M	3/10→M
Summarization	2/10→M	1/10→M 1/10→O	3/10→M
Education	1/10→M	6/10→M	3/10→M 1/10→O
Care plan	1/10→M	3/10→M	6/10→M
Other (O)	7/15→M	3/10→M	6/10→M

Table 2: Cluster relabeling between rounds. Elements of the table correspond to how many clusters of a given semantic class were re-labeled as (O)ther or (M)ixed.

Clustering sentence-level model. To cluster sentence-level embeddings, we use a similar approach to the one described in turn-level clustering (§ 5.2). The only difference is that we use the predicted labels to constrain that the kmeans++ algorithm is independently applied to examples corresponding to each predicted label. As an example, Figure B.1 in the Appendix shows the visualization of clusters predicted to be part of "Summarization." Each cluster is manually assigned its label (often simply staying with the original predicted label) by examining about ten data points (sentences).

Details of relabeling between rounds. Table 2 shows the number of clusters relabeled and the new label assigned. We can see that most relabeling was moving clusters to the "Mixed" label, thereby ensuring that we improve the 'purity' of the pseudo-labels. Examples with the "Mixed" label are not used for the subsequent round of model training. However, they would still be used for subsequent clustering and relabeling. This strategy of relabeling also helps to mix high-confidence predictions with low-confidence ones, as long as they are close in the embedding representations.

5.4. Implementation details

All models discussed are trained in Pytorch 1.10.2+cu102 with the language models implemented using HuggingFace Transformers library (Wolf et al., 2019). The weights for the DeCLUTR models were using the JOHNGIORGI/DECLUTR-BASE checkpoint. For training, we used the Adam optimizer with learning rate $2e^{-5}$ and a scheduler with warm-up steps of total training steps/5. We set the batch size as 12. PCA and kmeans were implemented using scikit-learn 0.24.2 package, while UMAP used the umap-learn 0.5.1 package.

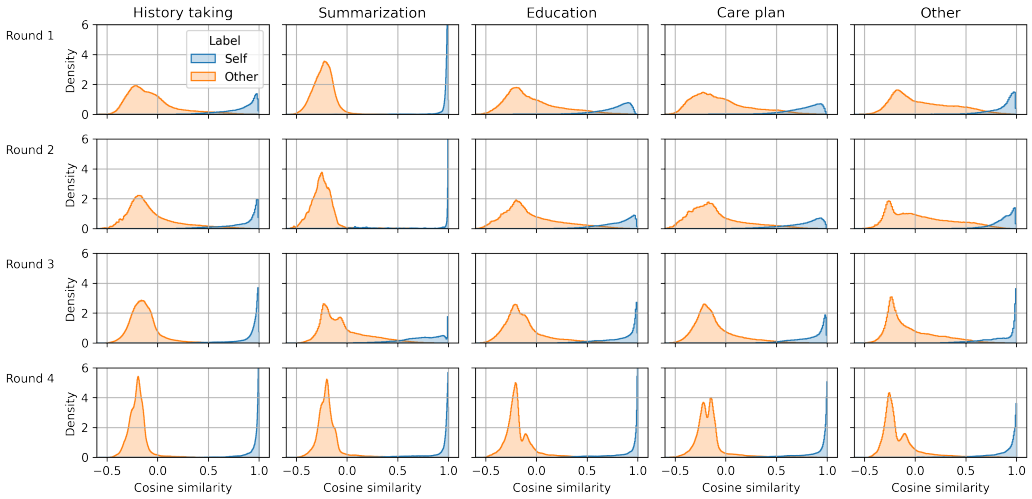

Figure 3: Cosine similarity of same- and different- class pairs for each class

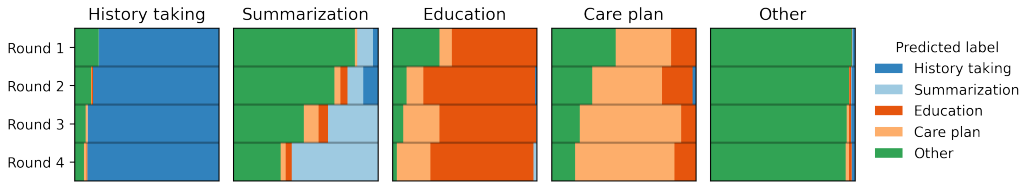

Figure 4: Improvement in label purity with each round (§ 6.1). Columns represent human-assigned true labels. Each row represents the proportion of the predicted labels as represented by the proportion of the color. The larger the proportion of the color corresponding to the column name, the better is the model improvement.

6. Results

6.1. Main result: Sentence-level model performance

Table 3 provides our main results, comparing F1 and accuracy scores from each training round of the sentence-level model. The overall performance increased from accuracy of 69.5% to 82.5%. The "Summarization" class has the most improvement (F1 score from 0.18 to 0.65). This three-fold improvement of the F1 score shows that our iterative approach can improve labeling quality (and hence the model) even when the initial labels are noisy. The sentences in this class are hard to identify solely from the turn-level-model-based pseudo-labeling. The pseudo-labeler successfully labels the sentences that contain "to summarize" but fails on *e.g.* "he experiences no pain." However, our iterative clustering-based labeling introduces less-confident predictions that are semantically similar to the more confident ones to improve the overall identifiability.

Figure 4 provides a graphical representation of the errors the model makes. Each column represents the human-assigned true label, and each row represents the proportion of the predicted labels in each true label for each training round. The two classes that see the F1 score uplift, "History taking" and "Summarization", start with a significant

	F1 score			
Class	**Round 1**	**Round 2**	**Round 3**	**Round 4**
Summarization	0.18±0.00	0.19±0.11	0.47±0.06	**0.65±0.02**
History Taking	0.89±0.01	0.9±0.00	0.92±0.00	**0.93±0.01**
Education	**0.70±0.01**	0.69±0.02	0.69±0.02	0.65±0.02
Care Plan	0.55±0.02	0.56±0.03	**0.57±0.01**	0.55±0.02
Other	0.90±0.00	0.92±0.00	**0.93±0.00**	**0.93±0.00**
Multi-class Accuracy*	69.5%±0.00	74.1%±0.00	80.4%±0.01	**82.5%±0.01**

* Accuracy is on the four functional classes only

Table 3: Sentence-level model performance: F1 scores and accuracy after each round of iterative training. Standard deviations by retraining models with different seeds.

		F1 score			
Class	**Turn-level**	**Round 1**	**Round 2**	**Round 3**	**Round 4**
Summarization	0.22±0.00	0.22±0.00	0.25±0.03	**0.69±0.04**	0.66±0.04
History Taking	0.37±0.00	0.84±0.02	0.83±0.01	0.86±0.01	**0.87±0.01**
Education	0.61±0.01	**0.77±0.02**	0.69±0.05	0.73±0.02	0.65±0.04
Care Plan	0.31±0.04	0.55±0.02	0.55±0.03	**0.57±0.01**	0.51±0.02
Other	0.75±0.00	0.89±0.01	0.93±0.00	**0.95±0.01**	**0.95±0.00**
Binary Accuracy	84.7%±0.00	**95.6%±0.00**	95.2%±0.01	**95.6%±0.00**	94.9%±0.00

* Accuracy is on the four functional classes only

Table 4: Turn-based inference improved with sentence-level model (§ 6.2). The column "Turn-level" is the initial turn-level model from which sentence level model was bootstrapped. Columns Round 1–4 show the F1-score when we pool sentence-level predictions to produce turn level labels. The standard deviations are derived by retraining models with different seeds.

confusion with the "Other" class, which gradually decreases. Even though the additional iterations did not improve the "Care plan" and "Education" classes, their overall confusion changed between rounds. Initially, both "Education" and "Care plan" were confused with the "Other" class, while in later rounds, they were confused with each other. We expect this inter-class confusion as they can be hard to differentiate even for human annotators, *e.g.* "It is recommended that a person having a fever should drink more water." could be annotated as either "Education" or "Care plan", depending on the context.

Figure 3 sheds light on another perspective on the change in the quality of the embeddings of the sentence-level models. Here, at every round, we randomly sampled 1,000 examples for each predicted class and used their embeddings to compute the distribution of cosine similarities between pairs of the same class ("self") and pairs of different classes ("other"). The distributions are always bimodal, but the full width at half max of the peaks decreases. Even for classes where the F1 metrics did not improve, there is an increase in the 'peakiness" of the two distributions, making them more separable. This is the separation between positive and negative contrastive learning examples, where recent literature on sen-

tence embeddings (Li et al., 2020; Liu et al., 2021) suggests that the increased separation corresponds to better generalization performance.

6.2. Can we obtain a better turn-level inference using the sentence-level model?

In the previous experiments, we evaluated the output of the sentence-level model for each sentence in the input. Here, we investigate if training models at the sentence level also improve turn-level performance. For this, we max pool the predictions of all the sentences in a turn. For comparison, we use the initial turn level model (§ 4.1) as the baseline.

Table 4 shows the F1 and accuracy scores of the sentence-aggregated turn-level predictions. Like the sentence-level models, we see the most marked improvement in the "Summarization" class. Note how the Round 1 sentence-level model outperforms the turn-level model even though the turn-level model is used to generate the sentence-level pseudo-labels at the beginning with no human relabeling. This suggests that the sentence-level model can learn better semantics that the turn-level model.

Overall, the improvement from the later rounds is less pronounced at the turn level. While sentence-level evaluation benefits from multiple rounds of disentangling the class confusion between sentences within a turn, this is less of a concern for turn-level evaluation. This is also evidenced by overall higher F1 scores when compared to evaluation at the sentence level in Table 3. The improved performance and decreased effect from additional iterations is likely because the sentences that are more difficult to classify into a particular class tend to appear in mixed class turns and therefore doing well on these sentences does not improve turn-level metrics.

7. Discussion

We proposed a method for automatically inferring functional sections of a patient-medical professional dialogue with minimal human supervisory data. While we focused on the four dominant medically relevant functional sections, "History taking", "Summarization," "Education," and "Care plan" along with a background ("Other") class, the approach can be easily extended to additional classes.

Starting with very little annotated data, we build a highly accurate model using a human-in-the-loop cluster-based pseudo-labeling strategy. We show that the approach increases embedding anisotropy, effectively increasing the contrast between labels. We think this is because our approach intermixes high and low-confidence predictions which are then relabeled on a per-cluster basis through a simple human-in-the-loop evaluation. This makes our label-refinement strategy potentially useful for other applications, where the starting pseudo-labels are noisy or insufficient to capture the data variability, and getting additional human labels is expensive.

Ethics This work was done as part of a quality improvement activity as defined in 45CFR §46.104(d)(4)(iii) – secondary research for which consent is not required for the purposes of "health care operations." All human annotators were full-time employees of the company while performing this work.

References

Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. In Abderrahim El Moataz, Driss Mammass, Alamin Mansouri, and Fathallah Nouboud, editors, *Image and Signal Processing*, pages 317–325, Cham, 2020. Springer International Publishing. ISBN 978-3-030-51935-3.

David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 978-0-898716-24-5.

Oleg Bestsennyy, Greg Gilbert, Alex Harris, and Jennifer Rost. Telehealth: A quarter-trillion-dollar post-covid-19 reality? https://tinyurl.com/2ead8992, 2021. [Online; accessed 31-March-2022].

Jiaao Chen, Zichao Yang, and Diyi Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *CoRR*, abs/2004.12239, 2020. URL https://arxiv.org/abs/2004.12239.

Caitlin Dreisbach, Theresa A Koleck, Philip E Bourne, and Suzanne Bakken. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International journal of medical informatics (Shannon, Ireland)*, 125:37–46, 2019. ISSN 1386-5056.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. Self-training improves pre-training for natural language understanding. *CoRR*, abs/2010.02194, 2020. URL https://arxiv.org/abs/2010.02194.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. 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 879–895, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.72. URL https://aclanthology.org/2021.acl-long.72.

Dong hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, 2013.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *EMNLP-Findings*, 2020.

Shafiq Rayhan Joty, Giuseppe Carenini, and Raymond T. Ng. Topic segmentation and labeling in asynchronous conversations. *CoRR*, abs/1402.0586, 2014.

Sopan Khosla, Shikhar Vashishth, Jill Lehman, and Carolyn Rose. Medfilter: Improving extraction of task-relevant utterances through integration of discourse structure and ontological knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7781–7797, 01 2020. doi: 10.18653/v1/2020.emnlp-main.626.

Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary Chase Lipton. Generating soap notes from doctor-patient conversations using modular summarization techniques. In *ACL*, 2021.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *EMNLP*, 2020.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *EMNLP 2021*, 2021.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Shubham Kumar Nigam, Angshuman Hazarika, Arnab Bhattacharya, and Ashutosh Modi. Semantic segmentation of legal documents via rhetorical roles. *ArXiv*, abs/2112.01836, 2021.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Sören Mindermann, Muhammed Razzak, Winnie Xu, Andreas Kirsch, Mrinank Sharma, Adrien Morisot, Aidan N Gomez, Sebastian Farquhar, Jan Brauner, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learned. *arXiv preprint arXiv:2107.02565*, 2021.

Ali Mottaghi, Prathusha K. Sarma, Xavier Amatriain, Serena Yeung, and Anitha Kannan. Medical symptom recognition from patient text: An active learning approach for long-tailed multilabel distributions. *CoRR*, abs/2011.06874, 2020.

Varun Nair, Namit Katariya, Xavier Amatriain, Ilya Valmianski, and Anitha Kannan. Adding more data does not always help: A study in medical conversation summarization with PEGASUS. *CoRR*, abs/2111.07564, 2021. URL https://arxiv.org/abs/2111.07564.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *CoRR*, abs/1804.09170, 2018.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. Soap notes. In *StatPearls*, page https://www.ncbi.nlm.nih.gov/books/NBK482263/. Treasure Island (FL): StatPearls Publishing, 2021. [Online; accessed 31-March-2022].

Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. User intent prediction in information-seeking conversations. *CoRR*, abs/1901.03489, 2019.

Vipul Raheja and Joel Tetreault. Dialogue Act Classification with Context-Aware Self-Attention. 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)*, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning, 2020.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S. Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *CoRR*, abs/2101.06329, 2021.

M. Saravanan, B. Ravindran, and S. Raman. Automatic identification of rhetorical roles using conditional random fields for legal document summarization. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008. URL https://aclanthology.org/I08-1063.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Ryuichi Takanobu, Minlie Huang, Zhongzhou Zhao, Fenglin Li, Haiqing Chen, Xiaoyan Zhu, and Liqiang Nie. A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, page 4403–4410, 2018.

Robert L. Thorndike. Who belongs in the family. *Psychometrika*, pages 267–276, 1953.

Ilya Valmianski, Nave Frost, Navdeep Sood, Yang Wang, Baodong Liu, James J. Zhu, Sunil Karumuri, Ian M. Finn, and Daniel S. Zisook. Smarttriage: A system for personalized

patient data capture, documentation generation, and decision support. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 75–96. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/valmianski21a.html.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Appendix A. Application of turn-level model on sentences

Input : Dialogue turns $\{T_j\}$
Sentences $S_{jk} \in T_j$
Universe of labels $\mathcal{L}$
Model $M_{\text{turn}}(T_j)$ for estimating section probability $P(L = L_i|T_j), L_i \in \mathcal{L}$
Output: Sentence level labels $L_{jk} \in \mathcal{L}$

$L_{\text{turn},j} \leftarrow \{l \in \mathcal{L} : P(L = l|T_j) > \alpha_1\}$
$L_{\text{filter},j} \leftarrow \{l \in \mathcal{L} : P(L = l|T_j) > \alpha_2\}$
foreach $S_{jk} \in T_j$ **do**
  if 'summarization' $\in L_{\text{turn},j}$ **then**
    $L_{jk} \leftarrow$ 'summarization'
  end
  else if $P(L =$ "history taking"$|S_{jk}) \geq \alpha_3$ **then**
    $L_{jk} \leftarrow$ 'history taking'
  end
  else if $P(L =$ "education"$|S_{jk}) \geq \alpha_3$ **then**
    $L_{jk} \leftarrow$ 'education'
  end
  else if $P(L =$ "care plan"$|S_{jk}) \geq \alpha_3$ **then**
    $L_{jk} \leftarrow$ 'care plan'
  end
  else
    $l_{\text{candidate}} \leftarrow \arg\max_c P(L = l|S_{jk}), l \in L_{\text{filter},j}$
    $L_{jk} \leftarrow l_{\text{candidate}}$ if $P(L = l_{\text{candidate}}|S_{jk}) > \alpha_1$ else "other"
  end
end
return L_{jk}

Algorithm 2: Pseudocode for applying the turn-level model to create sentence-level labels

Where $\alpha_1 = 0.5$, $\alpha_2 = 0.1$, and $\alpha_3 = 0.9$. The values were determined by an informal human evaluation of the pseudolabeling performance.

Appendix B. Example of Clustering outputs

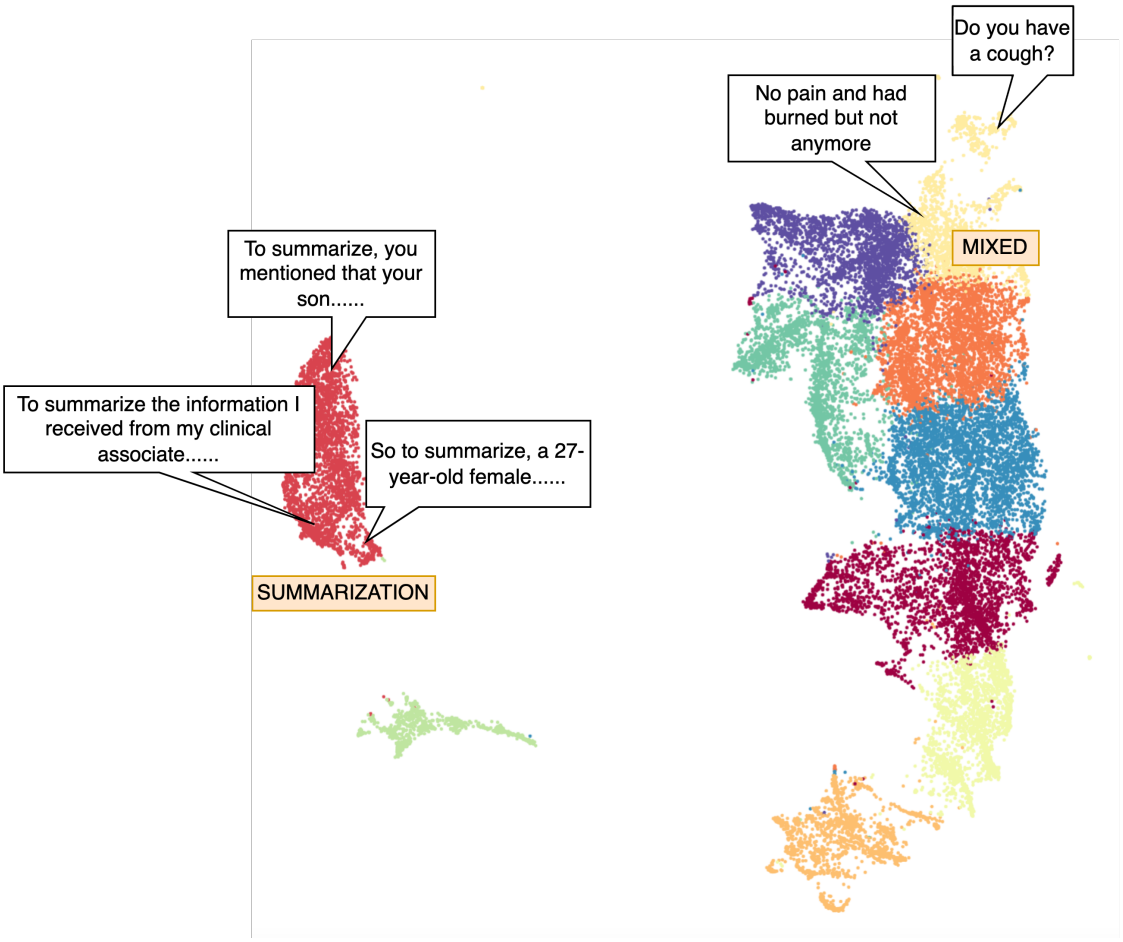

Figure B.1: Clustering of Sentences Predicted as "Summarization" after the First Round of Training

Appendix C. Examples of Encounters with Color-coded Model-generated Predictions

As our model is trained on predicting professionals' sentences, only the professionals' sentences are color-coded here.

Summarization

History Taking

Education

Care Plan

[MEDICAL PROFESSIONAL]
Hi ##. Thank you for completing our introductory questionnaire and for waiting few minutes we appreciate your time . My name is ## , and I am your Health Coach . My role allows me to educate you on your symptoms and provide additional resources regarding the topics we are able to address . I am not authorized to diagnose or prescribe . Over the next 20 minutes , I will gather some additional information so I can get a better picture of what is going on . What pill were you on ?
I understand you are worried about your concern and I am happy to help . Sorry for the delay earlier but it has been extremely busy today . I will try my best to be as quick as possible . I hope you understand and thank you for your patience

[PATIENT]
The combination pill

[MEDICAL PROFESSIONAL]
Since when are you on thus ?
This*

[PATIENT]
December 6-December 13

[MEDICAL PROFESSIONAL]
Alright , so if there was period following sex then the chances of pregnancy is slim

[PATIENT]
Okay even if Im 4 days late ?

[MEDICAL PROFESSIONAL]
Generally 3-7 days delays in cycle is considered normal. Anything beyond this needed to be evaluated in person by a obgyn

[PATIENT]
Okay I just thought I was super fertile from just having a baby
And I had unprotected sex the same day I took the pill for the first time

[MEDICAL PROFESSIONAL]
I understand , it would be wise to wait

[PATIENT]
Thank you

[MEDICAL PROFESSIONAL]
welcome

Figure C.1: Sample Encounter 1

[MEDICAL PROFESSIONAL]
Hi ## . Thank you for waiting we appreciate your time . My name is ## , and I am your Health Coach . My role allows me to educate you on your symptoms and provide additional resources regarding the topics we are able to address . I am not authorized to diagnose or prescribe . Over the next 20 minutes , I will gather some additional information so I can get a better picture of what is going on
Apologies for the delayed response . It has been quite busy today . Hope you understand
For how long has she been experiencing low sex drive ?

[PATIENT]
About 6 months

[MEDICAL PROFESSIONAL]
What do you mean by being sober ? . Was she addicted to any alcohol or drugs ?

[PATIENT]
Yes alcohol
She drank every day for about 15 years then her liver shut down then she quit drinking almost 2 years ago

[MEDICAL PROFESSIONAL]
Any history of known medical conditions or regular medications ?

[PATIENT]
No

Figure C.2: Sample Encounter 2

[MEDICAL PROFESSIONAL]
Alright . To answer your question , yes , the above symptoms can be associated with flu or allergies . It may take a few more days for the symptoms to subside . A few remedies like rest for the voice , staying well hydrated , warm salt water gargles , using a cool air humidifier or steam to prevent dryness of the throat , taking over the counter sore throat sprays and lozenges may help to ease some discomfort . Warm liquids , peppermint tea , warm water with honey can be soothing in most people . Generally , over the counter painkillers such as Tylenol or Ibuprofen for pain relief may help in many . Taking over the couter dry cough medicine may help too
Do you think you is able to try that ?

[PATIENT]
Yes , I can definitely do that . My throat is not sore though which is weird . I only cough when I feel what seems to be mucus in throat , but dont really spit up anything .

[MEDICAL PROFESSIONAL]
I understand . However , the above discussed remedies are known to help breakdown the mucus in the throat in most people and provide relief . :
Sounds good ?

[PATIENT]
That sounda good . So the chest tightness is mucus buildup ? . \
Sounds

[MEDICAL PROFESSIONAL]
Yes , you are right . Another possible cause can be narrowing of airways when they come in contact with allergen
Hope that answers your question

[PATIENT]
Yes it does . It makes my anxiety go up when I feel my chest do that , but good to know its nothing to worry about .
Thank you for the info

[MEDICALPROFESSIONAL]
You are most welcome . Please feel free to reach out with any questions in the future

[PATIENT]
Will do , ## .

[MEDICAL PROFESSIONAL]
Take care .

[PATIENT]
You too .

Figure C.3: Sample Encounter 3

[MEDICAL PROFESSIONAL]
Hi ## .Thank you for waiting we appreciate your time . My name is ## , and I am your Health Coach . My role allows me to educate you on your symptoms and provide additional resources regarding the topics we are able to address . I am not authorized to diagnose or prescribe . Over the next 20 minutes , I will gather some additional information so I can get a better picture of what is going on
How long have you had the symptoms ?

[PATIENT]
since late last night

[MEDICAL PROFESSIONAL]
Any fever ?

[PATIENT]
yes

[MEDICAL PROFESSIONAL]
Was the fever high grade or low grade ?

[PATIENT]
102

[MEDICAL PROFESSIONAL]
Any cough or cold ?

[PATIENT]
i cough off and on and if you mean am i cold yes

[MEDICAL PROFESSIONAL]
Yes
Do you have enlarged tonsils ?

[PATIENT]
from what someone felt they said they did feel big

[MEDICAL PROFESSIONAL]
Okay . Any white spots on the back of the throat ?

[PATIENT]
idk
i dont know

[MEDICAL PROFESSIONAL]
Do you think you can send a picture of the throat ?

[PATIENT]
im on a computer right now
give me a min

[MEDICAL PROFESSIONAL]
Okay .Take your time

[PATIENT]
i cant get one

[MEDICAL PROFESSIONAL]
Okay . Thats alright
Do you have any pain in abdomen ?

[PATIENT]
goes off and on

[MEDICAL PROFESSIONAL]
Where exactly is the pain located ?
How would you rate the pain on a scale of 1-10 , with 1 being minimum and 10 being maximum ?

[PATIENT]
its like a sharp pain in left lower and 5

[MEDICAL PROFESSIONAL]
Okay . Do you have any constipation ?

[PATIENT]
no

[MEDICAL PROFESSIONAL]
Okay . Have you tried any medications or measures to help with the symptoms ?

[PATIENT]
yes tyonle even ice

[MEDICAL PROFESSIONAL]
Okay . How long have you had the pain in the lower left abdomen ?

[PATIENT]
1 day

[MEDICAL PROFESSIONAL]
Okay . What else is happening ? . Are there any other symptoms apart from the pain such as nausea or vomiting ?

[PATIENT]
none of that
i have a headache

[MEDICAL PROFESSIONAL]
Where exactly is the pain in the head located ?

[PATIENT]
left back side a front right side

[MEDICAL PROFESSIONAL]
Okay . Did the headache also start yesterday or do you generally get headaches ?

[PATIENT]
i dont it slowly creeped up

[MEDICAL PROFESSIONAL]
Okay . Do you have any facial pain ?

[PATIENT]
no just the headache

[MEDICAL PROFESSIONAL]
Okay . To summarize our discussion so far , you have been having difficulty speaking and swallowing since yesterday along with sore throat and enlarged tonsils . Fever was 102 and Tylenol and ice were used . There is pain in left lower abdomen since yesterday . There is cough , cold and generalized body pain and fatigue as well along with a headache . There is no nausea / vomiting or constipation . Am I correct ?

[PATIENT]
yes correct

[MEDICAL PROFESSIONAL]
People with similar symptoms often experience an upper respiratory tract infection . Various causes such as tonsillitis , pharyngitis or strep throat or influenza can be present sometimes . If fever persists for more than 2 days or if symptoms worsen , it would be necessary to get checked in person

[PATIENT]
ok thank you

[MEDICAL PROFESSIONAL]
Youre welcome . Please let me know if you have any questions . I would like to check back in after a couple of days to see how you are doing . Thank you for letting me serve you . We are always here if you need anything

Figure C.4: Sample Encounter 4