Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2036–2044
Marseille, 11–16 May 2020

The Medical Scribe: Corpus Development and Model Performance Analyses

Izhak Shafran, Nan Du, Linh Tran, Amanda Perry, Lauren Keyes, Mark Knichel, Ashley Domin, Lei Huang,
Yuhui Chen, Gang Li, Mingqiu Wang, Laurent El Shafey, Hagen Soltau, Justin S. Paul
Google Inc.

Abstract

There is a growing interest in creating tools to assist in clinical note generation using the audio of provider-patient encounters. Motivated by this goal and with the help of providers and medical scribes, we developed an annotation scheme to extract relevant clinical concepts. We used this annotation scheme to label a corpus of about 6k clinical encounters. This was used to train a state-of-the-art tagging model. We report ontologies, labeling results, model performances, and detailed analyses of the results. Our results show that the entities related to medications can be extracted with a relatively high accuracy of 0.90 F-score, followed by symptoms at 0.72 F-score, and conditions at 0.57 F-score. In our task, we not only identify where the symptoms are mentioned but also map them to canonical forms as they appear in the clinical notes. Of the different types of errors, in about 19-38% of the cases, we find that the model output was correct, and about 17-32% of the errors do not impact the clinical note. Taken together, the models developed in this work are more useful than the F-scores reflect, making it a promising approach for practical applications.

Keywords: Medical Scribe, Information Extraction, Clinical Encounters, Span-Attribute Tagging (SAT) Model

1. Introduction

Medical providers across the United States are required to document clinical visits in the Electronic Health Records. This need for documentation takes up a disproportionate amount of their time and attention, resulting in provider burnout (Wachter and Goldsmith, 2018; Xu, 2018). One study found that full-time primary care physicians spent about 4.5 hours of an 11-hour workday interacting with the clinical documentation systems, yet were still unable to finish their documentation and had to spend an additional 1.4 hours after normal clinical hours (Arndt et al., 2017).

Speech and natural language processing are now sufficiently mature that there is considerable interest, both in academia and industry, to investigate how these technologies can be exploited to simplify the task of documentation, and to allow providers to dedicate more time to patients. While domain-specific automatic speech recognition (ASR) systems that allow providers to dictate notes have been around for a while, recent work has begun to address the challenges associated with generating clinical notes directly from speech recordings. This includes inducing topic structure from conversation data, extracting relevant information, and clinical summary generation (Quiroz et al., 2019). In one recent work, authors outlined an end-to-end system; however, the details were scant without empirical evaluations of their building blocks (Finley et al., 2018a). One of the simplistic approaches uses a hand crafted finite state machine based grammar to locate clinical entities in the ASR transcripts and map them to canonical clinical terms (Happe et al., 2003). This seems to perform well in a narrowly scoped task. A more ambitious approach mapped ASR transcripts to clinical notes by adopting a machine translation approach (Finley et al., 2018b). However this performed poorly. To address the difficulty in accessing clinical data, researchers have experimented with synthetic data to develop a system for documenting nurse-initiated telephone conversations for congestive heart failure patients who are undergoing telemonitoring after they have been discharged from the hospital (Liu et al., 2019). In their task, a question-answer based model achieved an F-score of 0.80. This naturally raises the question of how well state-of-art techniques will perform in helping the broader population of clinicians such as primary care providers.

One might expect that the task of extracting clinical concepts from audio faces challenges similar to the domain of unstructured clinical texts. In that domain, one of the earliest public-domain tasks is the *i2b2 relations challenge*, defined on a small corpus of written discharge summaries consisting of 394 reports for training, 477 for test, and 877 for evaluation (Uzuner et al., 2011). Given the small amount of training data, not surprisingly, a disproportionately large number of teams fielded rule-based systems. Conditional random field-based (CRF) systems (Sutton and McCallum, 2011) however did better even with the limited amount of training data (Uzuner et al., 2010). Other i2b2/n2c2 challenges focused on coreference resolution (Uzuner et al., 2012), temporal relation extraction (Sun et al., 2013), drug event extraction (Henry et al., 2019) on medical records, and extracting family history (Azab et al., 2019). Even though the text was largely unstructured, they benefited from punctuation and capitalization, section headings and other cues in written domain which are unavailable in audio to the same extent.

With the goal of creating an automated medical scribe, we broke down the task into modular components, including ASR and speaker diarization which are described elsewhere (Shafey et al., 2019). In this work, we investigate the task of extracting relevant clinical concepts from transcripts. Our key contributions include: (i) defining three tasks – the Medications Task, the Symptoms Task, and the Conditions Task along with principles employed in developing the annotation guidelines for them (Section 2.); (ii) measuring the label quality using inter-labeler agreements and refining the quality iteratively (Section 3.), (iii) evaluating the performance of the state-of-the-art models on these tasks (Section 4.), and (iv) a comprehensive analysis of the performance of the models including manual error categorization (Section 5.). The corpus we have created in this

work is based on private, proprietary data that cannot be publicly shared. Instead, we are sharing the learnings from our experience that might be useful for the wider community as well as the detailed labeling guidelines as supplementary material in the extended version of this paper on arxiv.org.

2. Corpus Development

The corpus of labeled conversations for this work was developed in conjunction with providers and medical scribes. In our first attempt, we annotated all the relevant clinical information with a comprehensive ontology that was designed for generating clinical notes using a slot-filling approach. This was found to be an extremely challenging task for the medical scribe labelers; it required a lot of cognitive processing, took a long time to label, and the results still had many discrepancies in quality and inter-labeler agreement.

2.1. Annotation Guidelines

The corpus described in this paper was annotated using certain guiding principles, described below.

First, the cognitive load on the labelers needs to be reasonably low to ensure high inter-labeler agreement. This consideration resulted in labeling each conversation in multiple passes, where the labelers focused only on a small subset of clinical concepts in each pass. They were instructed to ignore clinical information unrelated to the task at hand.

Second, the annotation of all the clinical information necessary to generate a clinical note was broken down into modular tasks, based on the type of entity, specifically, the symptoms, the medications, and the conditions. In each task, labelers focused on annotating the entities along with their coreferences and their attributes. For example, in the Medications Task, the labelers focused on the medications and their associated frequencies, dosages, quantities, and duration. The relationships between them were marked using undirected links. For simplicity, we ignored attributes of attributes, for example, taking one dosage of a medication in the morning and a different dosage in the evening.

Third, the guidelines were refined using a few iterations of experimentation and feedback to improve inter-labeler agreement before labeling the task. Finally, the ontology was pruned to retain clinical concepts with high inter-labeler agreement and sufficiently high occurrences so they could be modeled with the available data.

2.2. Description of the Corpus

The corpus consists of recordings and transcripts of primary care and internal medicine provider-patient encounters. The recordings were split into train, development and test sets consisting of about 5500, 500 and 500 encounters respectively. For measuring the generalization of the model results, the development and test splits were created in such a way that the providers are mutually exclusive. There were no identifiers associated with patients, thus there may be patient overlap across the sets.

2.3. Symptoms Task

The Symptoms Task focused on extracting symptoms described in the medical encounter so that they could be documented in the History of Present Illness (HPI), Review of Symptoms (ROS) and other sections of the clinical note. Thus, the ontology was designed to mirror the language and organization of symptoms in the note; the colloquial phrasing that patients use to describe symptoms is tagged with a clinically appropriate entity from the ontology.

Organ System	Symptom Entities
Constitutional	Const:Fever Const:Chills Const:Difficulty Sleeping
Gastrointestinal	GI:Abdominal Distension GI:Abdominal Pain GI:Vomiting
Neurologic	Neuro:Headache Neuro:Dizziness Neuro:Seizure

Table 1: An excerpt of the entities in the symptoms task.

As shown by an excerpt in Table 1, the ontology is organized in a categorical fashion, starting with a coarse-grained *organ system* which is further broken down into finer-grained *symptom entity tags*. The benefit of this approach is that the organ system can be used to organize symptoms into the ROS section of the note, even if the granular symptom entity isn't correctly inferred. The full ontology contains 186 symptom entities mapped to 14 organ systems.

The attributes for the Symptoms Task, illustrated in Table 2, were chosen based on the HPI elements in the CMS Evaluation and Management Service Documentation Guidelines (HHS.gov, 2017). After the manual labels were generated, the set was pruned to remove entities with low counts and low inter-labeler agreements as these would be hard to model. This resulted in 88 symptom labels.

Type	Attribute
Status	Experienced Not Experienced
Property	Duration Location Severity/Amount Frequency

Table 2: The attributes of symptom entities.

The status was marked by adding a second label to the entity. This choice sidestepped the more difficult task of identifying the words that signal whether a symptom was experienced or not. Unlike written text, the status is not explicitly mentioned. Instead, in a conversation, this is implied from the context, spread over multiple speaker turns and not easily associated with specific words.

The example in Table 3 illustrates how a conversation was labeled using the ontology for the symptoms task. There were a few challenges in accurately labeling the relevant content words. The annotators were instructed to assign the most specific label, but there were times when the context does not provide sufficient information. In such cases, they

PT: I've been having stomach issues around here for the last 2 weeks. It's bad.
DR: Okay, in the upper abdomen. What does it feel like?
PT: It kind of comes and goes and hurts. Sometimes I feel queasy.

Content Span	Symptom Label
stomach issues	GI:Other; Experienced
2 weeks	Property:Duration
bad	Property:Severity/Amount
upper abdomen	Property:Location
comes and goes	Property:Frequency
hurts	GI:Abdominal Pain; Experienced
queasy	GI:Nausea; Status:Experienced
sometimes	Property:Frequency

Table 3: An example symptoms task with its labels.

are instructed to assign the coarser system category along with a default of *Other*. Following this guideline, "*stomach issues*" in the above example would be assigned *GI:Other*.

In addition, we mapped patients' layman description of a symptom to a normalized form. In the illustrative example, the patient's described as feeling "*queasy*" and that is mapped to the clinical term *Nausea*. One advantage of this mapping is that we avoid an explicit normalization step that is commonly used in the literature.

Lastly, in provider-patient conversations, gestures and words are often used to indicate information about the symptoms, such as using the word "*here*" to refer to the part of the body where the symptom is experienced. The lack of verbalization of this information makes it difficult for an automated system to fully capture all relevant information. However, when providers clarify the location ("*upper abdomen*"), this helped in the ultimate task to automate the completion of the note.

2.4. Medications Task

The ontology for the Medications Task was designed to capture information related to all medications ranging from a patient's history to future prescriptions. While the symptoms could be reduced to a closed set (based on how they appear in the clinical notes), the list of medications is large and continually expanding. Therefore, the medication entity was treated as an open set, and the catch all *Drug* label was applied to all direct and indirect references to drugs. This included specific and non-specific references, such as "*Tylenol*", "*the pain medication*", and "*the medication*". The attributes related to the medication entity that were annotated include *Prop:Frequency*, *Prop:Dosage*, *Prop:Mode*, and *Prop:Duration*.

In the illustrative example in Table 4, the entities – "*diabetes medications*", "*Sulfonylurea*", "*Amaryl*", "*glimepiride*" – are annotated with the *Drug* label, while "*1 mg*", "*pill*", "*everyday*" are their attributes.

One challenge encountered in this task was choosing a specific label when other labels are equally valid. For example, "*90 day sample*" could be marked as the total quantity

DR: Are you taking any diabetes medication?
PT: My kidney doc just changed the pill.
DR: Oh, a Sulfonylurea. Like Amaryl?
The generic name is glimepiride.
PT: Yup, she started me on 1mg everyday.
DR: Do you use Insulin?
PT: The shot? Only my brother has to.

Content Span	Medications Label
diabetes medication	Drug
Sulfonylurea	Drug
Amaryl	Drug
glimepiride	Drug
pill	Property:Mode
1mg	Property:Dose
everyday	Property:Frequency
insulin	Drug
the shot	Property:Mode

Table 4: An example medications task with its labels.

consumed, or as the duration of the drug treatment. For increasing consistency, the labelers were asked to choose the labels using the preference in the order – *Dose, Frequency, Quantity, Duration, Mode*.

Additionally, patients often referred to their medication using vague descriptors, such as "*the pink pill*". In many cases, the providers are able to infer the medication from the context and so the annotators are instructed to label such mentions.

2.5. Conditions Task

The Conditions Task, like the other two tasks, was designed based on how the discussed condition shows up in the clinical note. Even though conditions have some overlap with symptoms, they refer to broader categories and are discussed typically using clinical terminology. Ambiguities on whether a mention is a symptom or a condition was resolved by relying on the ICD-10 Code database. The condition entities were categorized into – Condition:Patient, Condition:Family History, and Condition:Other. The attributes of conditions mirror that of symptoms, listed in Table 2, plus an additional tag to capture the onset of the condition, *Prop:Onset/Diagnosis*. An example of this task is illustrated in Table 5.

One challenge in the Conditions Task is that there were only a few mentions of conditions in each conversation, causing labelers to overlook the mentions inadvertently. This was mitigated with an automated method to improve recall as discussed in Section 3.2..

2.6. Relations

Once all entities and attributes have been tagged and extracted from a conversation, the attributes need to be linked to their associated entities in order for them to be useful in downstream applications such as filling out sections of a clinical note.

In this example, the duration and location are associated with "*stomach issues*" while the frequency attribute is as-

DR: Any history of diabetes?
PT: I have diabetes.
DR: When was that diagnosed?
PT: 10 years ago.
DR: OK, and it seems to be well-controlled.
Any history of high blood pressure in the family?
PT: My brother has early onset high blood pressure.

Content Span	Conditions Label
diabetes	Condition:Patient
10 years ago	Property:Time of Diagnosis
well-controlled	Property:Severity
high blood pressure	Condition:Family History
early onset	Property:Time of diagnosis

Table 5: An example conditions task with its labels.

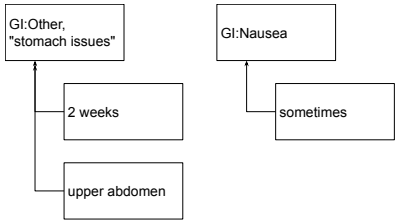

Figure 1: Illustration of relations for the example on the Symptoms Task.

sociated with "*nausea*". This would allow us to complete the following information in the note:

> Patient experienced stomach issues for 2 weeks in the upper abdomen. Patient also experienced nausea sometimes.

One difficulty we encountered in specifying links between entities and their attributes arose because of synonymous mentions. In the example illustrated in Table 4, the "*1 mg*" "*pill*" is equally related to "*Sulfonylurea*", "*Amaryl*" and "*glimepiride*". The annotators sometimes constructed an edge between the attributes and each of the drugs, and at other times chose one drug to be a canonical reference and create co-references for the other mentions of the drug. This makes it difficult to estimate the model performance reliably. The models developed for extracting relations using these annotations are described elsewhere (Du et al., 2019b).

3. Task Iteration

3.1. Training & Quality Assurance

Before starting a large scale annotation task, a number of steps were performed to refine the guidelines.
The process consisted of the following steps:

1. A small set of 3-5 conversations were labeled by a team of experienced labelers using the guidelines.
2. Differences in their labels were resolved by them in an adjudication session. In case the differences were large, steps (1) and (2) were repeated. This resulted in finalizing the guidelines and creating a reference set for instructional purposes.
3. The guidelines were then introduced to the larger team of labelers who were instructed to perform the same labeling task on the reference set.
4. Labelers were then scored based on agreement with the reference set. Those who scored well below the average quality score were given further training before proceeding to the task.

Additionally, in order to maintain consistently high quality labels while labeling the training set, we developed a quality assurance process. The process consisted of the following steps:

1. A team of experienced labelers created a reference set of labels for 3-5 conversations of varying complexity.
2. The rest of labelers were assigned the conversations from the reference set.
3. Labelers who scored highly with respect to the reference were chosen as reviewers.
4. The reviewers then reviewed the output of the other labelers by correcting the labels and documenting errors.
5. The feedback was sent to the labelers to incorporate into future labeling efforts.

Ultimately, we found that performing this K-way quality assurance process helped improve consistency significantly.

3.2. Automated Improvements

We explored different techniques to improve label quality. During the Conditions Task, we noticed that labelers often overlooked the mention of conditions because they occur too infrequently in the conversation. Since the conditions are often mentioned in their canonical form, we utilized the Google Knowledge Graph to identify them in the transcripts (Blog, 2012). The conditions identified by the knowledge graph were presented to the labelers as optional annotations that they can choose to use or discard. In order to minimize biasing the labelers, these optional annotations were presented only after the labelers finished their task and were in the process of submitting their conversation. In a controlled experiment, we found a 0.10 absolute improvement in recall from the labelers. Note, to avoid any potential measurement bias, this assistance was only provided while labeling the training set and not on the development and test sets where we relied labels from multiple labelers to maintain consistency as described further in Section 5.. Standard syntactic parsers were also investigated to pre-fill numberical values correspodning to dose, duration, frequency and quantity but was not employed in the labeling process.
We also experimented with using previously trained models to try to highlight numerical or date/time attributes, such as *Dose*, *Duration*, *Frequency*, or *Quantity*.

3.3. Common Challenges

While there were unique challenges for each of the tasks, a few challenges were common across them.

Despite the refinement of the ontologies and the continual training of labelers, there were still errors in labeling because the task is non-trivial and requires considerable attention. Certain residual errors were flagged using task-specific validation rules. These rules caught the most egregious errors, such as an attribute not being grouped with an entity, or not double-tagging a symptom or condition with its status.

A shared difficulty across all tasks was striking a balance between making the annotations clinically useful (i.e. developing a path to use the annotations to construct a complete HPI) while also reducing the cognitive burden for annotators. Through multiple experiments we uncovered significant challenges that could or could not be resolved with the addition, deletion, or substitution of tags in our ontologies.

The order in which clinically relevant events are described is often important in constructing a comprehensible HPI; for example, chest pain followed by shortness of breath and lightheadedness has different implications than shortness of breath followed by lightheadedness and chest pain. However, capturing this level of detail required expanding the ontology which was found to be cognitively burdensome for labelers and was left out to be addressed in future work.

Another challenge we encountered was confusion between temporal tags – *Time of Onset* or *Duration*, or progression information (e.g. *Improving* or *Worsening*). The words used to describe these attributes were often used in casual conversation manner. In the early version of the ontologies, the temporal information was annotated for each task differently, such as *SymProp:Frequency* and *MedsProp:Frequency*. This created confusion among the annotators. Subsequently, the temporal tags were defined uniformly across the tasks to the extent possible.

Another limitation of the annotation scheme is that they do not currently capture co-occurrence relationships between multiple entities although they might be clinically relevant. For example, "*nausea*" that a patient may experience along with "*abdominal pain*". The annotations also do not capture cross-ontology relationships, such as when a medication could be an alleviating factor for symptom or condition. This could be facilitated by showing labels from previous tasks when labeling a new task.

Even after carefully preparing before launching any labeling task, the labelers encountered novel situations that were not considered while developing the ontologies and the guidelines. This led to further refinements of the tasks including changing or adding a new tag. As a result, the portions of the data had to re-labeled, which was expensive. Analysis of the existing labels were used to guide these decisions, such as ignoring tags that appeared too infrequently, focusing on tags where there was a high level of disagreement, and looking at the distribution of labeled text for each tag.

While written domain corpora such as *i2b2* and *n2c2* also face similar challenges, the difficulty of this task is compounded by targeting a comprehensive ontology and information being spread across multiple speaker turns with large variations in how medical concepts are described by patients in the conversation (Uzuner et al., 2011; Henry et al., 2019).

3.4. Distributions & Inter-labeler Agreements

The distribution of the occurrences of the labels in the corpus are reported in Figure 2. Among entities, the number of medications mentioned is the highest. There are more attributes of symptoms than other tasks. Not surprisingly, the number of mentions of conditions and their attributes are the least among the three concepts.

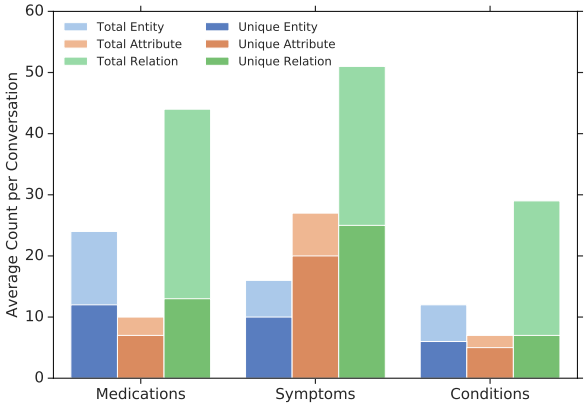

Figure 2: Numbers of labels and relations per conversation in three tasks. The unique counts are the number of unique occurrences accounting for the same highlighted span of text.

Clinical concepts vary in complexity considerably and in order to gauge the difficulty of the task and the quality of the labels, we estimated the inter-labeler agreements for the different tasks. The inter-labeler agreement was computed in terms of Cohen's kappa over the dev set of 500 conversations using three labelers per conversation. For scoring, we used a strict notion of a match that required both the label and the text content to be identical. In the Figure 3, we can see the inter-labeler agreements are higher for entities compared to attributes or relations. Among the three tasks, the medication achieves the highest inter-labeler agreement, followed by conditions and then symptoms.

The analysis of the inter-labeler agreement helped us identify poor labels and in certain cases modifying the names of the labels to be descriptive helped improve consistency. For example, replacing *Response* with *Improving/Worsening/Unchanged* was found to be useful.

4. The Span-Attribute Tagging Model

One of the challenges for model development in this task is the limited amount of training data. To use the data efficiently, we developed the Span-Attribute Tagging (SAT) model which performs inference in a hierarchical manner by identifying the clinically relevant span using a BIO scheme and classifying the spans into specific labels (Du et al., 2019a). The span is represented using the latent representation from a bidirectional encoder, and as such has the capacity to capture the relevant context information. This

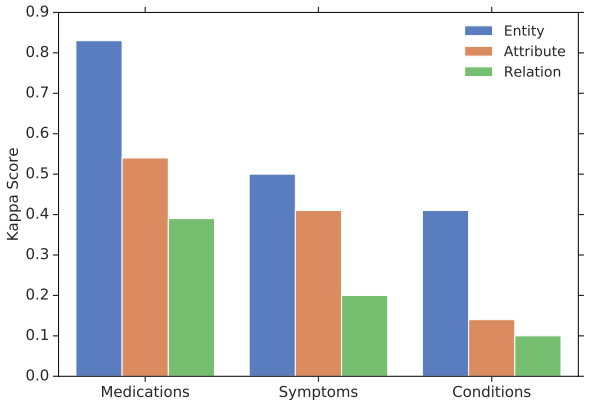

Figure 3: The inter-labeler agreement for entities, attributes, relations across three tasks.

latent contextual representation is used to infer the entity label and the status of the symptom as experienced or not. This is a trainable mechanism to infer status in contrast to *ad hoc* methods applied in negation tasks in previous work in clinical domain (Huang and Lowe, 2007).

For extracting relations, we extended it to Relation-SAT model which uses a two stage process: in the first stage the candidate entities are stored in a memory buffer along with its latent contextual representation and in the second stage a classifier tests whether the candidate attributes are related to the entries in the memory buffer (Du et al., 2019b). In both cases the models are trained end-to-end using a weighted sum of the losses corresponding to each inferred quantity.

5. Performance Evaluations

5.1. Voted Reference

For a robust evaluation of the models, we created a single "voted" reference from the 3-way labels mentioned in Section 3.4..

Directly applying a standard voting algorithm is not particularly useful in improving the reference. Consider the BIO labels from three labelers, shown in Table 6, for the phrase "*The pain medicine*". While all 3 labelers agree that a part of the phrase should be labeled, they have chosen different spans. A direct voting on the BIO annotations results in assigning "O" to the first token, and randomly choosing the second and third tokens.

	The	*pain*	*medication*
Labeler #1	Drug_B	Drug_I	O
Labeler #2	O	Drug_B	Drug_I
Labeler #3	O	O	Drug_B

Table 6: An example to illustrate the "voting" algorithm.

We address this issue using a Markov chain, which assigns the task tags separately from the BIO notation. For each time point t, let $Y^1(t)$ denote the tag assigned (e.g. *Drug*, *Frequency*, etc.) and $Y^2(t)$ denote the corresponding BIO notation (e.g. B, I, O). Each token can be represented by the tuple $(Y^1(t), Y^2(t))$. We explicitly define the state at each t as $Y^1(t-1)$ (with $Y^1(-1)$ defined as the empty set) and determine $Y^2(t)$ by picking the tags that maximize $P(Y^2(t)|Y^1(t), Y^1(t-1))$, where the probabilities are estimated empirically from the data. Applying this to the example above results in the annotation (,O), (*Drug*, B), and (*Drug*,I), which is a more reasonable outcome than the naive voting method.

5.2. Relaxed F-score

The annotators themselves are not always consistent in the span of text they label. E.g., "*pain medication*" vs. "*the pain medication*". To allow a certain degree of tolerance in the extracted span, the model performance was evaluated using a weighted F-score.

Let $Y_i(t)$ denote the parsed ground truth labels such that i corresponds to the label index and t corresponds to the token index within the sentence. We define $R_i = 1/|Y_i| \sum_{Y_i} I(Y_i(t) = \hat{Y}^1(t))$ to represent the label specific recall score for label i, where $\hat{Y}^1(t)$ represents the model predicted tag for time point t. The overall recall is estimated as $R = 1/n \sum_{i=1}^{n} R_i$ where n represents the total number of ground truth labels. Similarly, label specific precision is estimated as $P_j = 1/|\hat{Y}_j| \sum_{\hat{Y}_j} I(\hat{Y}_j(t) = Y^1(t))$, where $\hat{Y}_j$ represents the jth constructed label predicted by the model, and the overall precision is estimated as $P = 1/m \sum_{j=1}^{m} P_j$, where m represents the total number of predicted labels. The F1 score is subsequently calculated using the standard formula $2 * P * R/(P + R)$. Note that updating the label specific recall and precision scores to be the cumulative product (rather than the average) results in the well known strict scores (as they then enforce that the entire span to match, rather than just a subset of the span).

5.3. Model Performance Analysis

The performance of the SAT model on the three tasks are reported in Table 7. Because the Symptoms task involved a closed set of target classes for the entities, we evaluate performance by computing metrics at the conversation level ignoring the duplicates (e.g., symptoms repeated with the same status). The attributes were modeled using a common separate model.

Performance F1 (Precision, Recall)		
	Entities	Entities+Status
Symptoms	**0.72** (0.78, 0.66)	**0.60** (0.65, 0.56)
Conditions	**0.57** (0.61, 0.54)	**0.52** (0.56, 0.49)
Medications	**0.90** (0.94, 0.87)	—

Table 7: The performance of the SAT model for entities in the Symptoms, the Conditions, and the Medications tasks.

5.4. Manual Error Categorization

The key aim of manual error analysis is to supplement automatic evaluation of model performance by providing greater insight into the types of errors made by the model, the clinical relevance of the model errors, and plausible reasons why an error may have occurred. This analysis also allowed us to tease apart errors easily fixed in future model iterations with those that will be more difficult to correct.

Based on expected model output and anomalies in our dataset, we defined two major categories – *ErrorCause* and *ErrorImpact*. The raters were asked to utilize the context to provide their best guess on why the model might have caused the error. If they were uncertain, they could mark the cause as unknown. Likewise, the *ErrorImpact* assesses the relevance the error would have if the extracted information were populated into a clinical note. To mark an error as '*relevant*', the scribe was asked to judge whether the error would actually go into the medical note. For each type of error – Deletion, Insertion, and Substitution – we asked the raters to attach the labels associated with *ErrorCause* and *ErrorImpact*. The subcategories associated with these two major categories are listed in Table 8. The list includes cases such as model being correct, the inferred span being incorrect, the failure of the model to use the contextual cues, errors that are associated with needing additional medical knowledge not evident from the context, and attributes that may not be related to any relevant entities.

Category	Subcategory
ErrorCause	Agree with model Incorrect span Ambiguous tag Irrelevant attribute Fail to use context Need clinical expertise Break in conversation flow Clinically equivalent No clear reason
ClinicalRelevance	Relevant Not relevant N/A

Table 8: Categories and subcategories of the model errors.

The errors were categorized by a group of about 10 medically trained raters, including physicians, physician assistants and professional medical scribes. Raters were trained through self review of guidelines, live instruction with guided annotation, and practice conversations. Results of the practice conversations were compared to a conversation that had been evaluated and adjudicated by a group of three individuals who had created the task guidelines. Raters were provided a side by side view of their submission compared to the adjudicated conversation. Those that showed major deviation from the adjudicated conversation were given additional guidance by task managers.

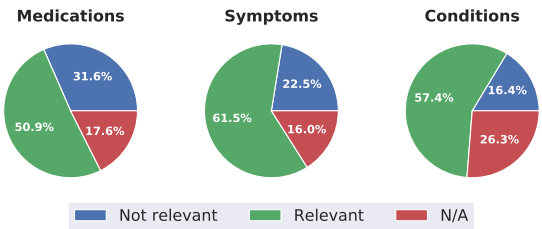

Figure 4: Proportion of errors relevant to the clinical note.

Among the errors committed by the model, about 17-32% of the errors do not impact the clinical note, as shown in Figure 4.

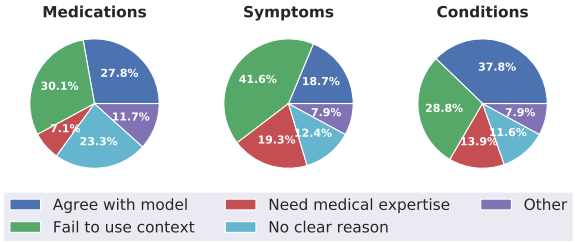

Figure 5: Proportion of errors attributed to causes.

Human raters found that the prediction from the models were correct in about 19-38% of the cases and that the errors were associated with the labels in the reference, as shown in Figure 5.

In analyzing the cases where the rater agreed with the model, we noticed that (despite using the voted reference labels) label quality issues still remained. Specifically, there was a noticeable amount of occurrences where the annotator missed or mis-labeled the documents. Taken together, the output of our model is significantly more useful in generating the clinical note than what is reflected in the F-score.

Among the other errors, in about 29-42% of the cases, the model did not take the context into account as much as humans do. For example, a patient might be describing the medication adherence as described in the example below. A human would understand that the doctor is referring to another medication that the patient is on. However, the model fails to capture this.

DR: Let's talk about your medications. How much of the Levemir are you using now?
PT: 60 units.
DR: And the other one?

About 7-19% of the errors relate to needing medical expertise beyond what is known from the context. In the example given below, the patient's description of their symptom ("*hard to breathe*") would be labeled as *Respiratory:Orthopnea*. However, the model picks *Respiratory:Shortness of Breath* which would be a reasonable choice but not sufficiently specific given the context ("*when I'm lying down*") and the medical knowledge relating to it. Future model improvements need to focus correcting these

PT: It's really hard to breathe. It's particularly hard when I'm lying down in bed.

two categories of errors.

6. Turn Detection Model

Our empirical results show that the model that we developed performs poorly on attributes compared to entities. One possible explanation is that unlike entities, the span boundaries of attributes may be less distinct in a conversation. We investigated this by relaxing the model requirements, from identifying the span of words to detecting

speaker turns containing the attributes, which happens to be similar to other previous work (Finley et al., 2018b; Lacson et al., 2006; Park et al., 2019).

The turn detection model we investigated consists of an encoder, similar to the SAT model, which is followed by an attention layer and a multi-label softmax. The output predicts the probability of the occurrence of the attribute in the given input turn.

In the experiment, six attributes were selected as the target classes: *Frequency*, *Duration*, *Location*, *Severity*, *Alleviating Factor*, and *Provoking Factor*. These attributes were selected based on higher frequency and inter-rater agreements. Also, the model was trained and evaluated on the labels merged across all three ontologies as well as for each ontology. The model performance is reported as the F1 score, precision, recall along with our SAT model in Figure 6.

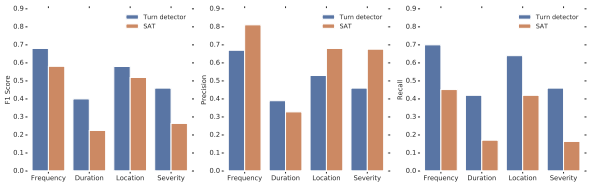

Figure 6: Comparison of performances on attributes.

The result shows that the turn-based detection approach achieves better recall (but lower precision) compared to our tagging-based SAT model. The trade-off shows that when the nature of the tags does not have distinct span boundaries, modeling them at the turn-level results in better performance, especially in the situation when recall is more crucial. Note, the turn detection model was trained by treating each speaker turn as an independent input. Clearly, this can be improved further by encoding the whole conversation and predicting the class labels for each turn, which should also improve the per task attribute score.

One potential application of the turn-based model is to assist the human annotation process. For example, the model predictions can be used to pre-select a set of turns that have high probability of containing targeted tags in order to help the annotators narrow down the scope to search. Or the model predictions can be used to compare against the annotators result to help capture potential tags the annotator missed. Based on our initial experiments, a combination of the above methods can improve the efficiency and quality of the human annotations. As aforementioned, such mechanism to improve labeling efficiency should not be applied to evaluation data to avoid introducing bias in performance measurement.

7. Discussion

To our knowledge this is the first systematic effort in extracting relevant information from provider-patient conversations with the aim of populating a clinical note. The informal nature of the conversation poses considerable challenges in extending standard methods of annotating a corpus, as described in Section 2..

The complexity of extracting the relevant clinical concepts from conversational speech is considerably higher than in the written domain. For example, in a comparable work, a standard CRF model was able to achieve an F-score of 0.9 in written clinical documents (Patel et al., 2018), whereas a simple CRF model performs poorly on our task (Du et al., 2019a). Similarly, while developing annotation guidelines, several examples presented considerable challenge in deciding how best to annotate them, a reflection of the difficulty of the task for humans and models.

One might argue that entities such as medication can be extracted by *ad hoc* approaches such as relying on dictionaries and regular expressions, as in (Finley et al., 2018b). However, such an approach would not generalize to common place references to medications such as "*the shot*" which in certain context may refer to insulin shot and in others may not be clinically relevant. In contrast, our approach generalizes to rare occurrences of the long-tailed distribution of medications. For example, our annotators captured 8,613 unique spans of medications in the documents reviewed.

Automatic methods to assist labelers proved to be useful. The labeling throughput can be further improved using the output of previously trained models, especially for entities and attributes where the model performance is sufficiently high. For example, the medications entities, the symptoms entities, the locations and the frequencies.

Results from our manual error categorization indicate that there is room for improving the annotations further. An example that we have successfully tested is implementing a quality assurance stage where randomly sampled conversations are reviewed by senior labelers and corrected prior to being submitted for modeling. Additionally, the conversations could be down-weighted in cases where the manual annotations are substantially different from cross-validated model predictions (Wang et al., 2019).

Another promising direction to improve the extraction of clinically relevant concepts would be to combine the SAT model and the turn detection model since they have complementary precision and recall trade-offs.

8. Conclusions

This paper describes a novel task for extracting clinical concepts from provider-patient conversations. We describe in detail the ontologies and the annotation guidelines for developing a corpus. Using this corpus, we trained a state-of-the-art Span-Attribute Tagging (SAT) model and report results that highlight the relative difficulties of the different tasks. Further through human error analyses of the errors, we provide insights into the weakness of the current models and opportunities to improve it. Our experiments and analyses demonstrate that several entities such as medications, symptoms, conditions and certain attributes can be extracted with sufficiently high accuracy to support practical applications and we hope our results will spur further research on this important topic.

9. Acknowledgements

This work would not have been possible without the support and help from a number of people, including Kyle Scholz, Nina Gonzalez, Chris Co, Mayank Mohta, Zoe Kendall, Roberto Santana, Philip Chung, Diana Jaunzeikare, and I-Ching Lee.

References

Arndt, B. G., Beasley, J. W., Watkinson, M. D., Temte, J. L., Tuan, W.-J., Sinsky, C. A., and Gilchrist, V. J. (2017). Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *The Annals of Family Medicine*, 15(5):419–426.

Azab, M., Dadian, S., Nastase, V., An, L., and Mihalcea, R. (2019). Towards extracting medical family history from natural language interactions: A new dataset and baselines. In *Proc. EMNLP-IJCNLP*, pages 1255–1260.

Blog, G. (2012). https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

Du, N., Chen, K., Kannan, A., Tran, L., Chen, Y., and Shafran, I. (2019a). Extracting symptoms and their status from clinical conversations. In *Proc. ACL*, pages 915–925, Florence, Italy, July. Association for Computational Linguistics.

Du, N., Wang, M., Tran, L., Li, G., and Shafran, I. (2019b). Learning to infer entities, properties and their relations from clinical conversations. In *Proc. EMNLP-IJCNLP*, pages 4980–4991, Hong Kong, China, November. Association for Computational Linguistics.

Finley, G., Edwards, E., Robinson, A., Brenndoerfer, M., Sadoughi, N., Fone, J., Axtmann, N., Miller, M., and Suendermann-Oeft, D. (2018a). An automated medical scribe for documenting clinical encounters. In *Proc. NAACL Demonstrations*, pages 11–15, New Orleans, Louisiana, June. Association for Computational Linguistics.

Finley, G., Salloum, W., Sadoughi, N., Edwards, E., Robinson, A., Axtmann, N., Brenndoerfer, M., Miller, M., and Suendermann-Oeft, D. (2018b). From dictations to clinical reports using machine translation. In *Proc. NAACL/HLT*, pages 121–128, New Orleans - Louisiana, June. Association for Computational Linguistics.

Happe, A., Pouliquen, B., Burgun, A., Cuggia, M., and Beux, P. L. (2003). Automatic concept extraction from spoken medical reports. *I. J. Medical Informatics*, 70(2-3):255–263.

Henry, S., Buchan, K., Filannino, M., Stubbs, A., and Uzuner, Ö. (2019). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*.

HHS.gov. (2017). Cms evaluation and management service documentation guidelines. `https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/eval-mgmt-serv-guide-ICN006764.pdf`. Accessed: 2019-11-26.

Huang, Y. and Lowe, H. J. (2007). A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *Journal of the American Medical Informatics Association*, 14(3):304–311, 05.

Lacson, R. C., Barzilay, R., and Long, W. J. (2006). Automatic analysis of medical dialogue in the home hemodialysis domain: Structure induction and summarization. *Journal of biomedical informatics*, 39(5):541–555.

Liu, Z., Lim, H., Suhaimi, N. F. A., Tong, S. C., Ong, S., Ng, A., Lee, S., Macdonald, M. R., Ramasamy, S., Krishnaswamy, P., Chow, W. L., and Chen, N. F. (2019). Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. In *Proc. NAACL (Industry Papers)*, pages 24–31, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Park, J., Kotzias, D., Kuo, P., Logan IV, R. L., Merced, K., Singh, S., Tanana, M., Karra Taniskidou, E., Lafata, J. E., Atkins, D. C., et al. (2019). Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. *Journal of the American Medical Informatics Association*, 26(12):1493–1504.

Patel, P., Davey, D., Panchal, V., and Pathak, P. (2018). Annotation of a large clinical entity corpus. In *Proc. EMNLP*, pages 2033–2042, Brussels, Belgium, October-November. Association for Computational Linguistics.

Quiroz, J. C., Laranjo, L., Kocaballi, A. B., Berkovsky, S., Rezazadegan, D., and Coiera, E. (2019). Challenges of developing a digital scribe to reduce clinical documentation burden. *Nature Partner Journals Digital Medicine*, 2.

Shafey, L. E., Soltau, H., and Shafran, I. (2019). Joint speech recognition and speaker diarization via sequence transduction. In *Proceedings of Interspeech*, volume abs/1907.05337, Graz, Austria, September.

Sun, W., Rumshisky, A., and Uzuner, Ö. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, September.

Sutton, C. and McCallum, A. (2011). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Uzuner, Ö., Solti, I., and Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, September.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Uzuner, Ö., Bodnari, A., Shen, S., Forbush, T., Pestian, J., and South, B. R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791, September.

Wachter, R. and Goldsmith, J. (2018). To combat physician burnout and improve care, fix the electronic health record. *Harvard Business Review*, March.

Wang, Z., Shang, J., Liu, L., Lu, L., Liu, J., and Han, J. (2019). Crossweigh: Training named entity tagger from imperfect annotations. In *Proc. EMNLP-IJCNLP*, pages 5153–5162, Hong Kong, China, November. Association for Computational Linguistics.

Xu, R. (2018). The burnout crisis in American medicine. *The Atlantic*, May.