Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 413–421
Marseille, 11–16 May 2020

Alignment Annotation for Clinic Visit Dialogue to Clinical Note Sentence Language Generation

Wen-wai Yim†, Meliha Yetisgen*, Jenny Huang‡, Micah Grossman†
† Augmedix, Inc., *University of Washington, ‡ University of California, Los Angeles
† 1161 Mission St 210, San Francisco 94103,
∗850 Republican St, Seattle WA 98109,
‡ 405 Hilgard Avenue, Los Angeles, CA 90095
wenwai.yim@augmedix.com, melihay@uw.edu, jyh08@ucla.edu, micah.grossman@augmedix.com

Abstract

For every patient's visit to a clinician, a clinical note is generated documenting their medical conversation, including complaints discussed, treatments, and medical plans. Despite advances in natural language processing, automating clinical note generation from a clinic visit conversation is a largely unexplored area of research. Due to the idiosyncrasies of the task, traditional methods of corpus creation are not effective enough approaches for this problem. In this paper, we present an annotation methodology that is content- and technique- agnostic while associating note sentences to sets of dialogue sentences. The sets can further be grouped with higher order tags to mark sets with related information. This direct linkage from input to output decouples the annotation from specific language understanding or generation strategies. Here we provide data statistics and qualitative analysis describing the unique annotation challenges. Given enough annotated data, such a resource would support multiple modeling methods including information extraction with template language generation, information retrieval type language generation, or sequence to sequence modeling.

Keywords: Corpus, Dialogue

1. Introduction

Two trends drive significant interest in automating the process of medical documentation: a growing shortage of clinicians in the United States and a rise in clinician burnout rates due to health information technology-related stress (Gidwani et al., 2017; AAMC, 2019; Gardner et al., 2019). Clinicians today are responsible for more than just the well-being of their patients. They must also complete documentation on their time present with the patient, diagnoses made, and treatments prescribed in the electronic medical record. Medical scribes, who can assist clinicians with completing medical documentation, are one way clinicians can unburden themselves from documentation responsibilities. But the cost of employing a medical scribe, estimated between $49K (onsite) and $23K (virtual) a year (Brady and Shariff, 2013), is prohibitive.

Despite recent natural language processing (NLP) advancements, such as improved speech to text, deep learning NLP modeling, and greater availability of clinical NLP resources, the task of converting a clinic visit conversation into its corresponding clinical note remains challenging. True comprehension of the clinical situation discussed in a visit requires many difficult aspects of language understanding and generation, such as summarizing over multiple statements and question-answer responses. Clinical note generation also depends on input from outside factors, e.g. electronic medical record data, templates, and patients' reported visit complaints. When a scribe is present, intake information can come from both a clinic visit conversation dialogue as well as direct communication with a medical scribe. This fluidity in dialogue and sourcing adds more complexity to the problem. Parts of the conversation may involve comforting a patient, clarifying information, or extracting information. What ultimately becomes documentation-worthy is often highly specialty-, institution-, and provider- specific. Properly trained NLP models will require large annotated corpora to achieve this ambitious goal and capture the ways non-sequitur statements or question-answer summarizations inform the final clinical note content. Previous annotation methodologies do not address the unique nature of this problem. Moreover, due to concerns over patient privacy, medical conversation data and clinical note data, each by themselves, are low-resource domains.

In this paper, we introduce a novel annotation methodology which matches clinical note sentences to grouped sets of dialogue sentences. The goal is to build a corpus of annotated data which can support a variety of techniques, including information extraction with template language generation or sequence to sequence methods. With such a corpus one could build systems to generate clinical notes from clinic visit audio automatically. This annotation approach flexibly adapts to the significant variability between provider, specialty, and institution. To our knowledge, this is the first work to attempt such a content- and technique- agnostic systematic annotation methodology for this task.

2. Background

While unique institutions and departments may have different practices, a pattern of common steps emerges in note creation. Pre-charting is the first step, done prior to the start of a visit: an appropriate note template is selected and populated where necessary with pertinent information related to the patient as well as reason for visit, indicated from scheduling. During the actual visit, the clinician converses with a patient to gather details of the problem, diagnose illnesses, and discusses treatments and plans. This is the data capture step. At times the clinician will prompt for medical context from the patient. At other times, the clinician col-

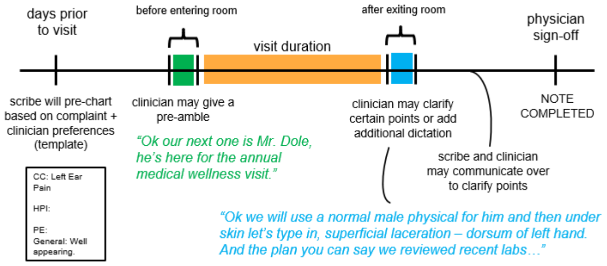

days prior to visit
before entering room
visit duration
after exiting room
physician sign-off
scribe will pre-chart based on complaint + clinician preferences (template)
CC: Left Ear Pain
HPI:
PE:
General: Well appearing.
clinician may give a pre-amble
“Ok our next one is Mr. Dole, he’s here for the annual medical wellness visit.”
clinician may clarify certain points or add additional dictation
scribe and clinician may communicate over to clarify points
NOTE COMPLETED
“Ok we will use a normal male physical for him and then under skin let’s type in, superficial laceration – dorsum of left hand. And the plan you can say we reviewed recent labs…”

Figure 1: Note creation life cycle with a scribe

Visit State	Note Sections
Discovering reason for visit / Verbal examination	History of Present Illness (HPI) Review of System (ROS) Social History (SHx) Routine Health Maintenance (RHM)
Physical examination	Physical Examination
Detailing treatment or further investigation	Assessment and Plan Impression

Table 1: Implied visit state and note section output dependencies

<table>
<tr><th>Note</th><th>Dialogue</th></tr>
<tr><td>She declines the pneumonia vaccine.</td><td>28 | <b>Doctor:</b> Have you had a pneumonia vaccine?
29 | <b>Patient:</b> No, I don’t think so.
30 | <b>Doctor:</b> Alright, do you want one?
31 | <b>Patient:</b> No.</td></tr>
</table>

Table 2: Summarization into note sentence across four turns of questions and answers

<table>
<tr><td>32 | <b>Doctor:</b> Ok, why don’t you lay down here so I can check your abdomen.
.. | ...
85 | <b>Patient:</b> Yeah I I also got a couple of moles here that I want you to check
86 | <b>Doctor:</b> Ok um you still trying to do exercise on top of work?
87 | <b>Patient:</b> Yeah I’ve been swimming.
89 | <b>Doctor:</b> Ok nice and symmetrical homogenous color.</td></tr>
</table>

Table 3: Intertwining visit states and anaphora

lects very specific information through focused questions. All significant information is captured in the clinical note during or shortly after the visit. The note is incomplete until the clinician formally signs off by adding their name to the note. If a scribe is employed, he or she may be responsible for these different aspects as well as with communicating with the clinician to clarify any uncertainty. In the case of a remote scribe, there may be extra steps such as a preamble where a clinician may describe the next patient or an after-visit clarification step. Figure 1 depicts a possible note creation cycle with a remote scribe working with a clinician.

Clinical notes are organized into note sections. Certain sections have implied intake procedures specific to certain parts of a normal visit. For example, “History of Present Illness” and “Review of Systems” sections cover the clinician verbally interviewing their patients; whereas the “Physical Examination” section is used when the clinician physically examines the patient; and finally “Assessment and Plan” section is populated with discussion of treatments or further investigations.

Table 1 shows example visit states and note section destinations. Table 2 shows the resulting note sentence from portions of a dialogue. Not only must information be extracted across two question answer adjacency pairs (and 4 turns), but the second question answer pair does not give an explicitly mentioned subject. The nature of conversations further complicates the matter, as multiple threads can happen at the same time covering different areas of the visit state. An example of this is shown in Table 3: during a physical exam, the clinician and patient may discuss other matters.

A full abbreviated dialogue exchange and its resulting clinical note are shown in Table 4 to illustrate such complexities. The note is organized into sections. Several noncontinuous parts of the dialogue may interweave (lines 84-

<table>
<tr><th>Note</th><th>Dialogue</th><th>Annotations</th></tr>
<tr><td>0 | Chief Complaint :
1 | Annual physical
2 | HPI :
.. | ...
.. | ...
5 | Requesting refills for medications
6 | He denies wheezing , nausea and vomiting .
.. | ...
.. | ...
18 | He swims and weight lifts for exercise
19 | He is married with two adult sons .
.. | ...
.. | ...
26 | Physical Exam
.. | ...
.. | ...
28 | Skin :
.. | ...
.. | ...
33 | Moles on abdomen are symmetrical, homogeneous in color , and non - raised .
.. | ...
.. | ...
62 | Assessment & Plan :
.. | ...
68 | Medications refilled .
.. | ...
.. | ...</td><td>0 | <b>Doctor:</b> The next patient is a 51 year old male presenting at the office today for his annual physical and follow up with this chronic problems of anxiety, hypertension, and hyperlipidemia.
.. | ...
17 | <b>Doctor:</b> And what about any problems breathing or wheezing?
18 | <b>Patient:</b> I feel like I can’t breathe as as deeply as I usually do.
.. | ...
35 | <b>Doctor:</b> No nausea, vomiting, diarrhea?
36 | <b>Patient:</b> Umm well diarrhea, yeah.
.. | ...
84 | <b>Doctor:</b> Ok, why don’t you lay down here so I can check your abdomen.
85 | <b>Patient:</b> Yeah I I also got a couple of moles here that I want you to check
86 | <b>Doctor:</b> Ok um you still trying to do exercise on top of work?
87 | <b>Patient:</b> Yeah I’ve been swimming.
88 | <b>Patient:</b> Doing some weight lifting.
89 | <b>Doctor:</b> Ok nice and symmetrical homogenous color.
.. | ...
118 | <b>Patient:</b> We usually go to Oregon to see our boys but they couldn’t come here and we couldn’t go so.
.. | ...
126 | <b>Doctor:</b> So where are your boys in Oregon?
127 | <b>Patient:</b> My youngest in is a nurse in Beaverton Washington County.
.. | ...
.. | ...
129 | <b>Patient:</b> My oldest is in Portland.
.. | ...
135 | <b>Patient:</b> Yeah and uh, my medications, I need refills.
136 | <b>Doctor:</b> Ok yep, we’ll have your meds refilled.</td><td>note[1] → STATEMENT2SCRIBE[0]
note[5] → STATEMENT[135]
note[6] → QA[17,18]
QA[35,36]
note[18] → GROUP
[QA[86,87],
QA[86,88]]
note[19] → GROUP
[STATEMENT[118],
QA[126,127],
QA[126,129]]
note[33] → GROUP
[STATEMENT[84],
STATEMENT[85],
STATEMENT[89]]
note[68] → STATEMENT[136]</td></tr>
</table>

Table 4: Example annotations (right) for corresponding clinical note (left) and dialogue (middle). The same colors indicate matched associations.

89), and anaphora[1] is ubiquitous (line 89). The order of appearances in note and dialogue often don't correspond, leading to many crossing annotations.

3. Related Work

While some work exists on doctor-patient conversation analysis (Byrne and Long, 1977; Raimbault et al., 1975; Drass, 1982; Cerny, 2007), annotation (Wang et al., 2018), and dialogue topic classification (Rajkomar et al., 2019), few explore the relationship between a patient visit's dialogue and a clinical note. We describe three groups with some coverage of the problem.

In (Kazi and Kahanda, 2019), the authors studied generating psychiatric case note sentences from doctor-patient conversation transcripts. Their work classified doctor-patient response pairs into semantic categories (e.g. client details, family history), then used a rule-based language processing system/tool/etc. to paraphrase the text into formal clinical text. Though an interesting idea, the data set was small (18 transcripts) and the authors do not assess the performance of their natural language generation.

In (Jeblee et al., 2019), the authors use 800 conversations to perform several classifications including: utterance type (e.g. question, statement), temporal and clinical entity extraction, attribute classification, classification of entities to a SOAP format and classification of primary diagnosis. Sentence generation was left for future work. This approach makes strong technique commitments, assumes a fixed clinical note template output, and does not lend well to support paraphrasing techniques.

In (Finley et al., 2018a), members of the EMR.AI team describe one intended approach to the problem by bridging information from clinic visit dialogue, first by classifying conversation sentences by intended note sections, then applying information extraction techniques. This data is then used to fill note templates generated by finite-state grammars. In another work (Finley et al., 2018b), they describe their method to automatically produce a parallel machine translation corpus for the special case of dictations to clinical note letter, but this focuses on just a narrow portion of the general problem.

We posit that the task of clinical note generation based on dialogue is best represented as an amalgamation of different language transformations and thus the annotation efforts should not be tied to specific end to end methods. Our proposed method associates a note sentence to associated dialogue sentence sets using different tags (e.g. DICTATION, QA, STATEMENT, etc) and provides a higher level ordering of these sets. Compared to (Kazi and Kahanda, 2019), (Jeblee et al., 2019), and (Finley et al., 2018a), we actually annotate the final note content output. Compared to the work in (Finley et al., 2018b), we manually create our alignments and do it for the entire conversation and note. Therefore our dataset does not rely on a specific sequence of domain-dependent NLU tasks or a specific relationship between the extracted information and the final output (e.g. template-filling), nor assumes a narrow part of the problem (e.g. dictation), freeing users to choose their own intermediate methods.

Our annotation objectives are inspired by and bear much similarity to the idea of machine translation corpus creation, the goal of which is to create sentence pairs which can be consumed by other algorithms (Koehn, 2005; Tiedemann, 2011). Several significant differences emerge from the end points being distinct mediums: one dialogue, one clinical note. For instance, dialogue data contains question and answer modes which must be mapped to prose. Additionally, the associations between the two mediums often occur out of sequence. Therefore, in contrast to machine translation corpus creation algorithms, our process cannot be easily automatized.

Our annotation methodology bears most similarity to that of (Hwang et al., 2015) and (Tian et al., 2014) who create parallel corpora by manually labeling paired sentences for aligned documents as good, good partial, partial, bad, etc., between Wikipedia and Simple Wikipedia and between Chinese-English online web articles, respectively. In contrast to their work, we distinguish between different types of dialogue to clinical note transformations, e.g. dictation, question-answering etc, as well as attempt to organize related sentences on the dialogue side into groups.

4. Data

The data comes from 66 mock patient visit materials used for training scribes. Though simulated, they were created to mimic real interactions. Material for each visit includes an audio recording and an associated clinical note. Audio recordings, 9.5 minutes duration on average with minimum, maximum and 50th-percentile at 2.1, 19.0, 9.5 minutes respectively, were run through Azure's speech to text service (azure.microsoft.com/en-us/services/cognitive-services/speech-to text, 2019). The transcript output was speaker-segmented manually and corrected for word errors. The dialogue transcript was sentence tokenized based on assigned punctuation. Notes were tokenized using spaCy (spacy.io, 2019). The total number of dialogue sentences was 11465, with a vocabulary of 4706 words, with averages of 95 number of turns[2] and 174 sentences per visit. The primary speakers were the clinician_primary, with 7133 sentences, 3135 turns; the patient, 4050 sentences, 2937 turns; in addition to several other speakers (Table 5). Clinical notes had a total of 3181 sentences, a mean of 48 sentences per visit, and a vocabulary of 2873 words.

Speaker	Sentences	Turns
clinician_primary	7133	3135
patient	4050	2937
guest_patientfamily	242	155
other	40	27

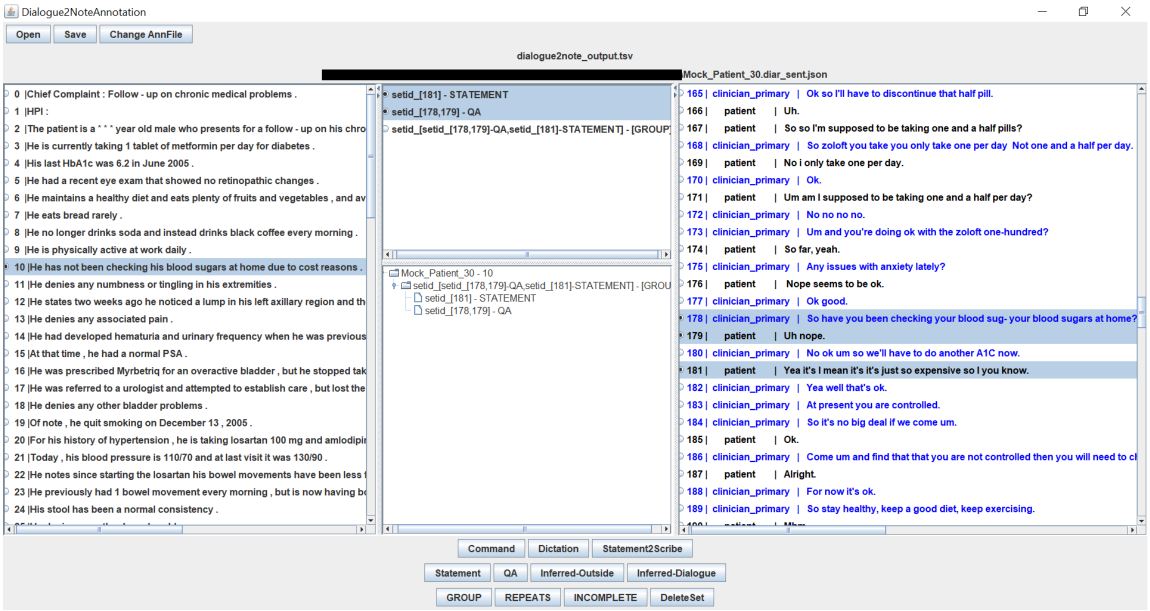
Table 5: Dialogue speaker breakdown

5. Annotation Methodology

The annotation methodology was created iteratively by a diverse group of clinical NLP experts, medical scribes, and

[1] Anaphora is the phenomenon when an expression can only be understood within context of another expression

[2] We define turn as a unit of continuous spoken language by one speaker before a speaker transition



dialogue2note_output.tsv

[redacted] Mock_Patient_30.diar_sent.json

```
0  |Chief Complaint : Follow - up on chronic medical problems .
1  |HPI :
2  |The patient is a * * * year old male who presents for a follow - up on his chro
3  |He is currently taking 1 tablet of metformin per day for diabetes .
4  |His last HbA1c was 6.2 in June 2005 .
5  |He had a recent eye exam that showed no retinopathic changes .
6  |He maintains a healthy diet and eats plenty of fruits and vegetables , and av
7  |He eats bread rarely .
8  |He no longer drinks soda and instead drinks black coffee every morning .
9  |He is physically active at work daily .
10 |He has not been checking his blood sugars at home due to cost reasons .
11 |He denies any numbness or tingling in his extremities .
12 |He states two weeks ago he noticed a lump in his left axillary region and th
13 |He denies any associated pain .
14 |He had developed hematuria and urinary frequency when he was previous
15 |At that time , he had a normal PSA .
16 |He was prescribed Myrbetriq for an overactive bladder , but he stopped tak
17 |He was referred to a urologist and attempted to establish care , but lost the
18 |He denies any other bladder problems .
19 |Of note , he quit smoking on December 13 , 2005 .
20 |For his history of hypertension , he is taking losartan 100 mg and amlodipi
21 |Today , his blood pressure is 110/70 and at last visit it was 130/90 .
22 |He notes since starting the losartan his bowel movements have been less
23 |He previously had 1 bowel movement every morning , but is now having b
24 |His stool has been a normal consistency .
```

```
setid_[181] - STATEMENT
setid_[178,179] - QA
setid_[setid_[178,179]-QA,setid_[181]-STATEMENT] - [GROUP]
```

```
Mock_Patient_30 - 10
  setid_[setid_[178,179]-QA,setid_[181]-STATEMENT] - [GROU
    setid_[181] - STATEMENT
    setid_[178,179] - QA
```

```
165 | clinician_primary | Ok so I'll have to discontinue that half pill.
166 |     patient       | Uh.
167 |     patient       | So so I'm supposed to be taking one and a half pills?
168 | clinician_primary | So zoloft you take you only take one per day  Not one and a half per day.
169 |     patient       | No i only take one per day.
170 | clinician_primary | Ok.
171 |     patient       | Um am I supposed to be taking one and a half per day?
172 | clinician_primary | No no no no.
173 | clinician_primary | Um and you're doing ok with the zoloft one-hundred?
174 |     patient       | So far, yeah.
175 | clinician_primary | Any issues with anxiety lately?
176 |     patient       | Nope seems to be ok.
177 | clinician_primary | Ok good.
178 | clinician_primary | So have you been checking your blood sug- your blood sugars at home?
179 |     patient       | Uh nope.
180 | clinician_primary | No ok um so we'll have to do another A1C now.
181 |     patient       | Yea it's I mean it's it's just so expensive so I you know.
182 | clinician_primary | Yea well that's ok.
183 | clinician_primary | At present you are controlled.
184 | clinician_primary | So it's no big deal if we come um.
185 |     patient       | Ok.
186 | clinician_primary | Come um and find that that you are not controlled then you will need to c
187 |     patient       | Alright.
188 | clinician_primary | For now it's ok.
189 | clinician_primary | So stay healthy, keep a good diet, keep exercising.
```

Figure 2: Annotation Software

a trained computer scientist with healthcare experience. The inter-annotator agreement was performed by the latter three. The corpus was annotated using an in-house software shown in Figure 2.

During annotation, each sentence of the clinical note can be attributed to multiple sets of sentences from the dialogue. Each labeled association of this type creates a set (In the software, to accomplish this, an annotator can select one note sentence and many dialogue lines and click a tag, e.g. STATEMENT, to create a new set which appears in the top middle panel). Each set should only include consecutive sentences of the same label, of the same speaker. The exception to this is the QA label which is expected to be 1 question, 1 answer, not necessarily consecutive. The QA label may also be used to mark a question without an answer. High level tags can only act upon other sets: for example, an annotator can only add a GROUP tag by selecting from the available previously created sets. The hierarchy is shown in the bottom middle panel. In the following sections, we describe the tags in further detail.

5.1. Annotations

Tags Addressing the Scribe

When the clinician is speaking outside of a conversation mode, we employ multiple tags to capture possible fundamental linguistic differences, such as when performing dictation or speaking to a scribe.

- COMMAND: Commands that specify document changes, which are attached to note section headers, e.g *"Use normal PE"*.
- DICTATION: Spoken information intended to be dictated, e.g. *"Mildly elevated liver enzymes, likely related to increased alcohol use the week prior to the labs"*.
- STATEMENT2SCRIBE: Other spoken information to scribe.

Template-related Tags

Clinical notes are often built on prefabricated templates. An annotator is assumed to have access to the originating note template and its default values. In cases when the note sentence comes from the template's default values, an annotator will apply one of the tags below.

- INFERRED-DIALOGUE: Non-explicit spoken information that can imply a template default (e.g. if in the dialogue it is known that a system is checked *"let me listen to your lungs"*) but there is no explicit mention of abnormality, we can infer that the default value is correct *"lungs: clear to auscultation bilaterally"*.
- INFERRED-OUTSIDE: Otherwise marks note sentences as coming from a template default.

Conversation Tags

During conversation, we identify two major modes of information exchange: question-answer and statements.[3]

- STATEMENT: Statements spoken during clinic visit conversation, e.g. *"She is here for her annual physical"*.
- QA: Question-answer modes during clinic visit conversation, e.g. *"Any vomitting or diarrhea? No."*.

Higher-level Tags

To mark some small amount of structure between normal sets, we use several higher level tags.

- GROUP: Used to group together discontinuous sets that are anaphoric.
- REPEATS: Used to indicate when a note sentence can be separately derived from different dialogue sets.

A clinical note's sentence can be left unattributed when its content is only derivable from outside knowledge, e.g. laboratory data. An already labeled note sentence may additionally be marked with an INCOMPLETE label when some information is unidentifiable from the dialogue.

To make the annotation task tractable, we implemented sev-

[3] This is consistent with, though a simplification of Todd's classification of doctor-patient communication speech acts : statements, questions, answers, directives, and reactives; for which we are not interested in the latter two (Cerny, 2007)

Note	Dialogue
He has also noticed mucous in his stool since last night.	1 | **Patient:** Oh I've I've had like diarrhea for like two days and then last night I was on the john and I thought I was just going to pass pas but mucous came out. 12 | **Doctor:** So you're saying you noticed when you had diarrhea you have mucus right? 13 | **Patient:** Yea diarrhea and then I had mucus.

Table 6: Salience rule example. Only the first sentence, which additionally indicates *"last night"*, should be associated in this case.

Note	Dialogue
She sustained muscle loss, and while her symptoms have gradually improved, she continues to have neuropathy.	139 | **Doctor:** Did you lose muscle then? 140 | **Patient:** Ohh yeah. 157 | **Doctor:** You've still lost a bit of muscle there didn't you? 158 | **Patient:** Mhmm.
No wheezes, rales or rhonchi	82 | **Doctor:** Hm lungs sound okay here. 110 | **Doctor:** And uh just a basic exam on him and all was normal.

Table 7: Repeat examples

eral high level annotation decisions. Mainly, when one annotation set contains more information related to the note sentence, other dialogue sentences with less information do not need to be annotated. The idea is to only annotate the most salient dialogue passages. An example of this is shown in Table 6. When multiple sets cover the same note sentence with equal salience, they are marked as REPEATS. Even if two dialogue sets are not repeating exactly the same information, the associations should be marked as repeats if we can derive the same note content. An example of this is shown in Table 7. Finally, note section headers are not annotated except for relevant COMMAND tags.

5.2. Inter-annotator Agreement

A single match between one note sentence and its associated dialogue sentences can be represented as a match tree, as shown in Figure 3. We evaluate matches according to three metrics: unlabeled triple, path, and span metrics. The first is a simple unlabeled f1 metric of selected dialogue sentences. An example from Figure 3 is 'mock_patient_01|note_18|86'. The second metric is an f1 measure where each instance is the full path from one note sentence to one dialogue sentence (e.g. 'mock_patient_01|note_18|GROUP|QA|86'). This metric is similar to the leaf-ancestor metric used in parsing, though we do not take path similarities. The final metric is a node-level labeled span of dialogue sentences, similar to that of PARSEVAL, e.g. An example of a span metric for the top group node would be 'mock_patient_01|note_18|GROUP|[86,87,88]' (Sampson and Babarczy, 2003). These reflect measures for simple matches as well as vertical and horizontal evaluations.

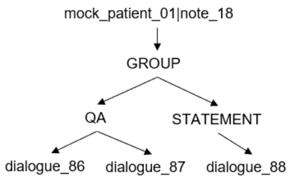


Figure 3: Annotation Match Tree

6. Quantitative Analysis

Table 8 shows the final inter-annotator agreement for a total of 10 clinic visits. Unsurprisingly, unlabeled triple had the highest agreement, with lower agreement for more complex metrics. Table 9 shows unlabeled triple agreements broken down by category. INFERRED-DIALOGUE tagged sentences were low as it requires judgement over what dialogue can infer the information from the note. We purposefully did not specify very specific guidelines for it, as both INFERRED and INFERRED-DIALOGUE are chiefly to mark parts of the note that came from templates and its default values.

	triple	path	span
A1/A2	0.73	0.40	0.61
A1/A3	0.78	0.53	0.66
A2/A3	0.69	0.36	0.57

Table 8: Agreements

label	A1/A2	A1/A3	A2/A3
COMMAND	0.77	0.75	0.55
DICTATION	0.65	0.72	0.69
INFERRED-DIALOGUE	0.06	0.26	0.07
INFERRED-OUTSIDE	0.84	0.67	0.60
QA	0.72	0.78	0.68
STATEMENT	0.57	0.70	0.57
STATEMENT2SCRIBE	0.38	0.68	0.39
INCOMPLETE	0.00	0.29	0.00

Table 9: Unlabeled match agreement breakdown

Out of all available note sentences, 81 ± 1 % were marked. In contrast, out of all dialogue lines 39 ± 11 % were marked. To generate clinical notes from dialogue, the note's author has to filter or aggregate significant amounts of information.

Max, min, median tree heights were 6, 4, 3 respectively. This supports our annotation design which encourages shallower trees, by only annotating the most salient evidence. For GROUP sets, 68% include 2 sets, 22%, 6%, and 2% for groups of 3, 4, and 5 sets respectively. For REPEAT sets, 92% contain two sets, 8% three sets.

The percentage of note and dialogue sentences with at least one label is shown in Table 10. In contrast, Table 11 shows the frequencies of a note sentences with the total number of associated tags. While note sentences with only STATEMENT dialogue information occurred the most frequently, this made up only 16% of all sentences, while DICTATION only 3%. Interestingly as high as 21% of note sentences required a GROUP label and that composition of multi-tagged note sentences were the most frequent. Together, this suggests that summarizing over multiple types of dialogue source information (e.g. STATEMENT and QA) is often required for note content generation.

We also measured the difference between the maximum and minimum dialogue lines for the top 3 frequent labels,

label	note	dialogue
COMMAND	0.8	0.1
DICTATION	3	1
GROUP	21	20
INFERRED-DIALOGUE	4	1
INFERRED-OUTSIDE	13	–
QA	28	17
STATEMENT	39	16
STATEMENT2SCRIBE	12	2
INCOMPLETE	3	–
REPEATS	4	3

Table 10: Percentage of sentences per label

Label-set	Freq	%	Cum. %
{STATEMENT}	496	16	16
{INFERRED-OUTSIDE}	395	12	28
{QA}	305	10	38
{QA,STATEMENT,GROUP}	279	9	46
{STATEMENT2SCRIBE}	238	7	54
{STATEMENT,GROUP}	189	6	60
{INFERRED-DIALOGUE}	106	3	63
{DICTATION}	98	3	66

Table 11: Top 8 occurring tagsets per note sentence. Column 4 is the sum percentage top to down.

which is informative for understanding how to group dialogue (Table 12). Typically, most occurrences of STATEMENT sets were single line (n=0), while the most frequent of QA was over two lines (n=1); though sometimes a question appeared without an answer (n=0). This shows that while most paired associations are a single sentence, there is spread of information across proximal sentences in a dialogue for STATEMENT. For QA, while many times the answer to a question can be found after the question, this is not always the case. Finally, the spread of required dialogue lines for GROUP label sentences (which account for 21% of sentences) suggests that to capture all related information per note sentence requires gathering related sentences spread across the dialogue.

n	STATEMENT	QA	GROUP
0	1546	73	0
1	253	1108	20
2	66	131	109
3	20	38	104
4	7	10	81
≥5	7	8	202

Table 12: Dialogue ranges (n lines) for associated top occurring labels per note sentence.

Analyzing the similarity of matched text between dialogue and note, we calculate the jaccard coefficient for unigrams[4] and UMLS concept identifiers, tagged with Metamap (Aronson and Lang, 2010), for the associated texts. The low similarity scores shown in (Table 13a) suggests that to get full matching context, simple similarity algorithms would be challenging. To quantify alignment difficulty, we calculate the percentages of sentences that cross n other sentences for tags exclusively not directed at the scribe, as well as for all sentences (Table 13b). For example, for n=3, we see that 76% of note sentences have evidence that crosses with at least three other note sentences. Since the preamble and after-visit clarifications may often provide salient information, if we did not count COMMAND, DICTATION, or STATEMENT2SCRIBE (the NON-SCRIBE) column, we would still find 60% of note sentences have information that crosses with at least three other note sentences' annotations. In all, the high percentages show that cross matches occurs frequently which would make automatic sentence alignment challenging.

percent.	concept	unigram
max	1.00	1.00
75	0.22	0.25
50	0.10	0.12
25	0.00	0.00

(a) Jaccard similarity

n	NON-SCRIBE	ALL
1	68 ± 17	85 ± 10
3	60 ± 22	76 ± 17
5	50 ± 24	66 ± 21

(b) % crossing annotations

Table 13: Data statistics of paired associations

7. Qualitative Analysis

In this section, we give qualitative descriptions along with examples to give the reader further insight into annotation disagreements. Below we describe several features of the annotation problem that present annotation challenges: (1) tag ambiguities related to the domain, (2) the problem of aggregating information and annotating over dialogues, and (3) the difficulty of using the GROUP label in the face of anaphoric references.

7.1. Domain-related Tag Ambiguities

On manual analysis, there were some common domain-related tag disagreement issues: (1) confusion between STATEMENT2SCRIBE vs DICTATION, (2) STATEMENT2SCRIBE vs STATEMENT, and (3) INFERRED-DIALOGUE and INFERRED-OUTSIDE.

For the most part when it is clear the clinician is speaking to the scribe and word-for-word translation is required, it is very apparent that associated label should be a DICTATION. However, there are some cases when there is a blend of speech for which the spoken information should be paraphrased and some in which direct copy would work. Though this is not a large problem (the agreement for DICTATION is amongst the highest) this is an interesting phenomenon. An example of this is shown in Table 14.

Confusion between STATEMENT2SCRIBE and STATEMENT occurs for cases in which a clinician can be interpreted as either speaking to the scribe or to the patient. This is more typical during the physical exam.

Note	Dialogue
Right maxillary sinus tenderness	[DICTATION] 48 | **Doctor:** Ok so there's right maxillary sinus tenderness.
Left frontal sinus tenderness	[STATEMENT2SCRIBE] 49 | **Doctor:** We have left sided frontal sinus tenderness.
Normal nasal mucosa	[STATEMENT2SCRIBE] 46 | **Doctor:** Ok, nasal mucosa is normal.

Table 14: Confusable STATEMENT2SCRIBE and DICTATION examples

[4] stop words and INFERRED-DIALOGUE lines were removed

The INFERRED-* tags denote parts of the clinical note that belong to default values of templates which is never explicitly mentioned in the dialogue. INFERRED-DIALOGUE is meant to capture parts of a dialogue that may suggest certain information. However, what is considered to be suggestive is somewhat subjective at times.

7.2. Annotating Over Lengthy Dialogue

A hallmark of our task is the requirement of the annotator to identify the most representative information of a clinical note sentence across an entire dialogue. This is especially difficult given the length of dialogues (on average 174 sentences per visit). This problem is compounded by the requirements of identifying the most salient information across many repeats of the same topics with various levels of information completion, as well as the unpredictable ordering of topics in the dialogue as compared to their appearance in the clinical note.

One alternative would include having annotators mark only first appearance. Another strategy would be to require annotators to mark every relevant sentence. In the former case, this strategy would forgo easy capture of variations within the same conversation – it would also forgo future abilities to automatically measure differences in expression of the same information within the same conversation. In the latter case, the amount of required annotations would vastly increase; furthermore the paired associated dialogue text would then contain much more repeating bits of information which would make the paired association less readily useful for artificial learning applications.

7.3. Grouping and Anaphora

Anaphora is ubiquitous in natural language, and especially in dialogue. The GROUP tag is used to mark such instances. For the special case in which referring expressions for pronouns and determiners need to be captured, the annotator is required to additionally mark the closest dialogue passage with the subject and connect the two sets with a GROUP tag. At times, this may be far from the actual information and may be ambiguous. An example is shown in Table 15. Different annotators may identify different passages to what constitutes an acceptable referential named entity. For cases in which additional anaphoric connections are required on top of the already identified referring expressions for pronouns and determiners, annotators – according to guidelines – are meant to connect to higher nodes. The exact hierarchy ordering can be easily perturbed amongst different annotators. Table 16 shows a complex group example.

Note	Dialogue
She reports the discharge has been worsening over the past two days.	setid_[2] - STATEMENT 2 | patient : Yes I have some discharge coming from my left eye. setid_[9] - STATEMENT 9 | patient : Well it's been feeling funny the past week waking up, but in the past couple of days I couldn't even open it when I woke up. <u>HIGHER LEVEL TAGS</u> setid_[setid_[2]-STATEMENT,setid_[9]-STATEMENT] - [GROUP] setid_[2] - STATEMENT setid_[9] - STATEMENT

Table 15: GROUP label used for pronoun and determiners

Note	Dialogue
Could be viral infection or a side effect of cephalexin, but also concerned regarding C. difficile infection.	setid_[38,39] - QA 38 | clinician_primary | Have you been on any antibiotics lately? 39 | patient: Uh Cephalexin. setid_[59] - STATEMENT 59 | clinician_primary: See you know any antibiotics can give you a little diarrhea. setid_[72] - STATEMENT 72 | clinician_primary: And also gram stain culture and test for c diff. setid_[128,129] - QA 128 | patient: So uh what what else could be causing this? 129 | clinician_primary: See it could also just be a virus causing all of this. setid_[130] - STATEMENT 130 | clinician_primary: So let's rule out the bad stuff but it could be a virus too. <u>HIGHER LEVEL TAGS</u> setid_[setid_[38,39]-QA,setid_[59]-STATEMENT] - [GROUP] setid_[38,39] - QA setid_[59] - STATEMENT setid_[setid_[128,129]-QA,setid_[130]-STATEMENT] - [REPEATS] setid_[130] - STATEMENT setid_[128,129] - QA setid_[setid_[72]-STATEMENT,setid_[setid_[128,129]-QA, setid_[130]-STATEMENT]-[REPEATS],setid_[setid_[38,39]-QA, setid_[59]-STATEMENT]-[GROUP]] - [GROUP] setid_[setid_[38,39]-QA,setid_[59]-STATEMENT] - [GROUP] setid_[72] - STATEMENT setid_[setid_[128,129]-QA,setid_[130]-STATEMENT] - [REPEATS]

Table 16: Complex group example

8. Discussion

Given the complexity and innate ambiguity of the task, we believe our agreements are good. Inconsistency between annotator's associations does not signify incorrectness (e.g. a sentence can have equally correct constituency parses). Table 17 shows three annotator markings for the same sentence in which all have annotation errors. The resulting annotation match trees are very different, but the content from each annotator markings are accurate. For example, A1/A2 triple, path, and span F1 agreements for this instances are 0.50, 0.13, 0.30; A1/A3, 0.77, 0.31, 0.54; A2/A3, 0.15, 0.00, 0.07. In comparison, the work by (Hwang et al., 2015) marking alignments between Simple Wikipedia and Wikipedia pages, which we believe to be closest to our task, achieved an annotator agreement of 0.68 Kappa for 46 articles and 67,853 sentence pairs. Our analogous agreement metric of simple match f1 was on average 0.73, which is consistent with this baseline.

In future iterations, we will add user-friendly improvements, including text search and highlights of clinically important elements to better assist annotating. Though sentences are the main unit of alignment here, this may be practically adjusted for real data: e.g. an automatically generated table in the note should be considered one unit instead of trying to divide the table into sentences, etc. While the data and analysis here was created from mock patient visits, the dialogue content should approximate actual real patient visits.

<table>
<tr><th>Annotator</th><th>Annotations</th><th>Errors</th></tr>
<tr><td>A1</td><td>setid_[] - [INCOMPLETE]
setid_[185] - STATEMENT
185 | patient : I eat yogurt.
setid_[187] - STATEMENT
187 | patient : I eat yogurt every day.
setid_[217] - QA
217 | clinician_primary : Are you taking any calcium?
setid_[224] - STATEMENT
224 | patient : Well I do drink milk.
setid_[230,231,232] - QA
230 | clinician_primary : Do you drink milk though, you know milk and cheese?
231 | patient : I eh yeah I mean I said I drink milk every day.
232 | patient : Cheese, yogurt every day.
HIGHER LEVEL TAGS
setid_[setid_[185]-STATEMENT,setid_[187]-STATEMENT,setid_[224]-STATEMENT,setid_[230,231,232]-QA] - [REPEATS]
setid_[185] - STATEMENT
setid_[230,231,232] - QA
setid_[187] - STATEMENT
setid_[224] - STATEMENT</td><td>Missing Vitamin D
QA should be 2 lines only
QA set mention of yogurt is more detailed therefore previous ones not required</td></tr>
<tr><td>A2</td><td>setid_[183,187] - QA
183 | clinician_primary : You taking probiotics or anything?
187 | patient : I eat yogurt every day.
setid_[183,185] - QA
183 | clinician_primary : You taking probiotics or anything?
185 | patient : I eat yogurt.
setid_[220] - STATEMENT
220 | clinician_primary : Yeah, so take at least twelve hundred of calcium and about two thousand of vitamin D.
setid_[222,223] - QA
222 | patient : So you want me to take the vitamin D even before we take the test?
223 | clinician_primary : Well, we will get the test, but I'll bet you're deficient.
setid_[224] - STATEMENT
224 | patient : Well I do drink milk.
setid_[232] - QA
232 | patient : Cheese, yogurt every day.
HIGHER LEVEL TAGS
setid_[setid_[183,185]-QA,setid_[183,187]-QA] - [REPEATS]
setid_[183,187] - QA
setid_[183,185] - QA
setid_[setid_[183,185]-QA,setid_[183,187]-QA,setid_[232]-QA] - [REPEATS]
setid_[232] - QA
setid_[183,187] - QA
setid_[183,185] - QA</td><td>Second mention of yogurt is more detailed therefore preivous one not required.
setid_[183,187] and setid_[183,185] cannot be considered repeats as first set supplies more complete information to the note sentence.
setid_[232] QA is missing the question
INCOMPLETE missing as we don't know the answer to whether calcium is taken.</td></tr>
<tr><td>A3</td><td>setid_[] - [INCOMPLETE]
setid_[217] - QA
217 | clinician_primary : Are you taking any calcium?
setid_[230,231] - QA
230 | clinician_primary : Do you drink milk though, you know milk and cheese?
231 | patient : I eh yeah I mean I said I drink milk every day.
setid_[230,232] - QA
230 | clinician_primary : Do you drink milk though, you know milk and cheese?
232 | patient : Cheese, yogurt every day.</td><td>Missing Vitamin D</td></tr>
</table>

Table 17: Different annotator's associations for the clinical note sentence *"She drinks milk and eats yogurt and cheese daily, but does not take calcium or vitamin D."*

9. Conclusions

In this work, we introduce a new annotation methodology for marking paired associations between a clinic visit dialogue and its complete clinical note. Given a large enough corpus of annotated data, the noise related to the difficulty of the task can be overcome to support multiple methods such as information extraction with templated generation, information retrieval type language generation, or sequence to sequence modeling. The same corpus can be used to train models for reordering generated sentences into proper required underlying clinical note structure or for building classifiers for slots in predetermined clinical note templates. In future work, we will apply this annotation approach for a large corpus of real patient visits to aid in clinical note content creation. Furthermore, we will train matching algorithms to assist annotation. Finally, after incorporating clinical note generated suggestions to aid our scribing operations, we can additionally use user-feedback for re-ranking output.

10. Acknowledgments

We would like to thank the Augmedix Scribe Operations who was responsible for source material creation and cleaning. We especially want to thank Jane Kwong for early contributions to the annotation schema creation and iteration.

11. Bibliographical References

AAMC. (2019). New findings confirm predictions on physician shortage.

Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–236.

azure.microsoft.com/en-us/services/cognitive-services/speech-to text. (2019). Speech to text API | microsoft azure.

Brady, K. and Shariff, A. (2013). Virtual medical scribes: making electronic medical records work for you. *J Med Pract Manage*, 29(2):133–6.

Byrne, J. F. and Long, P. (1977). Doctors talking to patients. *Psychological Medicine*, 7(4):735.

Cerny, M. (2007). On the function of speech acts in doctor-patient communication. *Linguistica*.

Drass, K. A. (1982). Negotiation and the structure of discourse in medical consultation. *Sociology of health& illness*.

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 *NAACL-HLT*.

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 *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 121–128, New Orleans - Louisiana, June. Association for Computational Linguistics.

Gardner, R., Cooper, E., Haskell, J., Harris, D., Poplau, S., Kroth, P., and Linzer, M. (2019). Physician stress and burnout: the impact of health information technology. *Jamia*, 26:106–114.

Gidwani, R., Nguyen, C., Kofoed, A., Carragee, C., Rydel, T., Nelligan, I., Sattler, A., Mahoney, M., and Lin, S. (2017). Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: A randomized controlled trial. *Annals of Family Medicine*, 15(5):427–433.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning sentences from standard wikipedia to simple wikipedia. In *HLT-NAACL*.

Jeblee, S., Khan Khattak, F., Crampton, N., Mamdani, M., and Rudzicz, F. (2019). Extracting relevant information from physician-patient dialogues for automated clinical note taking. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 65–74. Association for Computational Linguistics.

Kazi, N. and Kahanda, I. (2019). Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 140–148, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation.

Raimbault, G., Cachin, O., Limal, J. M., Eliacheff, C., and Rappaport, R. (1975). Aspects of communication between patients and doctors: an analysis of the discourse in medical interviews.

Rajkomar, A., Kannan, A., Chen, K., Vardoulakis, L., Chou, K., Cui, C., and Dean, J. (2019). Automatically charting symptoms from patient-physician conversations using machine learning. 179(6):836–838.

Sampson, G. and Babarczy, A. (2003). A test of the leaf-ancestor metric for parse accuracy | natural language engineering | cambridge core. 9(4):365–380.

spacy.io. (2019). spaCy Â· industrial-strength natural language processing in python.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., and Yi, L. (2014). UM-corpus: A large english-chinese parallel corpus for statistical machine translation. In *LREC*.

Tiedemann, J. (2011). Bitext alignment. *Synthesis Lectures on Human Language Technologies*, 4(2):1–165.

Wang, N., Song, Y., and Xia, F. (2018). Constructing a Chinese medical conversation corpus annotated with conversational structures and actions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).