Towards Automating Medical Scribing : Clinic Visit Dialogue2Note Sentence Alignment and Snippet Summarization

Wen-wai Yim
Augmedix Inc
wenwai.yim@augmedix.com

Meliha Yetisgen
University of Washington
melihay@uw.edu

Abstract

Medical conversations from patient visits are routinely summarized into clinical notes for documentation of clinical care. The automatic creation of clinical note is particularly challenging given that it requires summarization over spoken language and multiple speaker turns; as well, clinical notes include highly technical semi-structured text. In this paper, we describe our corpus creation method and baseline systems for two NLP tasks, clinical dialogue2note sentence alignment and clinical dialogue2note snippet summarization. These two systems, as well as other models created from such a corpus, may be incorporated as parts of an overall end-to-end clinical note generation system.

1 Introduction

As a side effect of widespread electronic medical record adoption spurred by the HITECH Act, clinicians have been burdened with increased documentation demands (Tran et al.). Thus for each visit with a patient, clinicians are required to input order entries and referrals; most importantly, they are charged with the creation of a clinical note. A clinical note summarizes the discussions and plans of a medical visit and ultimately serves as a clinical communication device, as well as a record used for billing and legal purposes. To combat physician burnout, some practices employ medical scribes to assist in documentation tasks. However, hiring such assistants to audit visits and to collaborate with medical staff for electronic medical record documentation completion is costly; thus there is great interest in creating technology to automatically generate clinical notes based on clinic visit conversations.

Not only does the task of clinical note creation from medical conversation dialogue include summarizing information over multiple speakers, often the clinical note document is created with clinician-provided templates; clinical notes are also often injected with structured information, e.g. labs. Finally, parts of clinical notes may be transcribed from dictations; or clinicians may issue commands to adjust changes in the text, e.g. "change the template", "nevermind disregard that."

note	dialogue
She declines the pneumonia vaccine.	[QA-1] Doctor: Have you had a pneumonia vaccine? [QA-1] Patient: No, I don't think so. [QA-2] Doctor: Alright, do you want one? [QA-2] Patient: No.

Table 1: Alignment example

In earlier work (Yim et al., 2020), we introduced a new annotation methodology that aligns clinic visit dialogue sentences to clinical note sentences with labels, thus creating sub-document granular snippet alignments between dialogue and clinical note pairs (e.g. Table 1, 2). In this paper, we extend this annotation work on a real corpus and provide the first baselines for clinic visit dialogue2note automatic sentence alignments. Much like machine translation (MT) bitext corpora alignment is instrumental to the progress in MT; we believe that dialogue2note sentence alignment will be a critical driver for AI assisted medical scribing. In the dialogue2note snippet summarization task, we provide our baselines for generating clinical note sentences from transcript snippets. Technology developed from these tasks, as well as other models generated from this annotation, can contribute as part of a larger framework that ingests automatic speech recognition (ASR) output from clinician-patient visits and generates clinical note text end-to-end (Quiroz et al., 2019).

2 Background

Table 2 depicts a full abbreviated clinical note with marked associated dialogue transcript sentences. To understand the challenges of alignment (creation of paired transcript-note input-output) and generation (creation of the note sentence from

Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, pages 10–20
July 6, 2021. ©2021 Association for Computational Linguistics

note	dialogue	annotations
0 | Chief Complaint : 1 | Evaluation of tonsil hypertrophy 2 | HPI : .. | ... .. | ... 5 | Reports enlarged tonsils, tonsil stones and sore throat. 6 | Symptoms have been present for several years but have worsened over the past several months. .. | ... 18 | She wakes up in the morning with nausea. 19 | She has frequent tonsil infections, 3-4 infections per year. .. | ... .. | ... 26 | Physical Exam .. | ... 28 | Turbinates : 29 | Normal size and symmetrical bilaterally. .. | ... .. | Tonsil : 33 | 3+ cryptic .. | ... .. | ... 62 | Assessment & Plan : .. | ... 68 | [Risk and benefits template for tonsillectomy] .. | ...	0 | **Doctor:** alright enlarged tonsils. .. | ... 6 | **Doctor:** okay so tell me about your throat. 7 | **Patient:** my tonsils they stay pretty big and they have tonsil stone and - .. | ... 9 | **Patient:** um like this once on this side specifically it's actually swollen- 10 | **Patient:** and a couple weeks ago it was so swollen that it was like bleeding. 11 | **Patient:** I wake up in the mornings and I feel like I'm going to be sick. .. | ... 18 | **Doctor:** so you had this for a long time? 19 | **Patient:** yeah 20 | **Doctor:** wait how old are you? 21 | **Patient:** twenty two. 22 | **Doctor:** and you've had tonsil infections since high school? 23 | **Patient:** mhm. .. | ... 24 | **Doctor:** sore throats? 26 | **Patient:** yeah. .. | ... 32 | **Patient:** do you think it happens more than three times in a year? 33 | **Patient:** probably at least three. .. | ... 48 | **Doctor:** tonsils three plus cryptic . .. | ... .. | ... 147 | **Doctor:** please insert the risks and benefits template for tonsillectomy.	note[1] → STATEMENT2SCRIBE[0] note[5] → GROUP [STATEMENT[6], STATEMENT[7], STATEMENT[9,10]] note[6] → GROUP [QA[18,19], QA[20,21], STATEMENT[22,23]] INCOMPLETE note[18] → STATEMENT[11] note[19] → QA[32,33] note[29] → INFFERRED-OUTSIDE note[33] → DICTATION[48] note[68] → COMMAND[147]

Table 2: Example annotations (right) for corresponding clinical note (left) and dialogue (middle). The same colors indicate matched associations.

the dialogue snippet), it is important to consider several differences in textual mediums:

Semantic variations between spoken dialogue and written clinical note narrative. Spoken language in clinic visits have vastly different representations than in highly technical clinical note reports. Dialogue may include frequent use of vernacular and verbal expressions, along with disfluencies, filler words, and false starts. In contrast, clinical note text is known to use semi-structured language, e.g. lists, and is known to have a much higher degree of nominalization. Moreover, notes frequently contain medical terminology, acronyms, and abbreviations, often with multiple word senses.

Information density and length. Whereas clinical notes are highly dense technical documents, conversation dialogue are much longer than clinical notes. In fact, in our data, dialogues were on average three times the note length. Key information in conversations are regularly interspersed.

Dialogue anaphora across multiple turns is pervasive. Anaphora is the phenomenon in which information can only be understood in conjunction with references to other expressions. Consider in the dialogue example : "Patient: I have been having swelling and pain in my knee. Doctor: How often does the knee bother you?" It's understood that the second reference of "knee" pertains to the knee-related swelling and pain. A more complex example is shown in Table 2 note line 6. While anaphora occurs in all naturally generated language, in conversation, it may appear across multiple turns many sentences apart with contextually inferred subjects.

Order of appearance between source and target are not consistent. The order of information and organization of data in a clinical note may not match the order of discussion in a clinic visit dialogue. This provides additional challenges in the alignment process. Table 2 shows corresponding note and dialogue information with the same color.

Content incongruency. Relationship-building is a critical aspect of clinician-patient visits. Therefore visit conversations may include discussion unrelated to patient health, e.g. politics and social events. Conversely, not all clinical note content necessarily corresponds to a dialogue content. Information may come from a clinical note template or various parts of the electronic medical record.

Clinical note creation from conversation amalgamates interweaving subtasks. Elements in a clinic visit conversation (or accompanying speech introduction) are intended to be recorded or acted upon in different ways. For example, some spoken language may be directly copied to the clinical note with minor pre-determined edits, such as in a dictation, e.g. "three plus cryptic" will be converted to "3+ cryptic". However some language is meant to express directives, pertaining to adjustments to the note, e.g. "please insert the risks and benefits template for tonsillectomy." Some information is meant to be interpreted, e.g. "the pe was all normal" would allow a note sentence "CV: normal rhythm" as well as "skin: intact, no lacerations".

Finally, there are different levels of abstractive summarization over multiple statements, questions and answers as shown in the Table 2 examples.

3 Related Work

Clinical Conversation Language Understanding Language understanding of clinical conversation can be traced to a plethora of historical work in conversation analysis regarding clinician-patient interactions (Byrne and Long, 1977; Raimbault et al., 1975; Drass, 1982; Cerny, 2007; Wang et al., 2018). More recent work has additionally included classification of dialogue utterances into semantic categories. Examples include classifying dialogue sentences into either the target SOAP section format or by using abstracted labels consistent with conversation analysis (Jeblee et al., 2019; Schloss and Konam, 2020; Wang et al., 2020). The work of (Lacson et al., 2006) framed identifying relevant parts of hemodialysis 118 nurse-patient phone conversations as an extractive summarization task. There has also been numerous works related to identifying topics, entities, attributes, and relations from clinic visit conversation – using various schemas (Jeblee et al., 2019; Rajkomar et al., 2019; Du et al., 2019). Though clinic conversation language understanding is not explored in this work, our automatic or manual sentence alignments methods produce the language understanding labels that may to used to (a) model dialogue relevance, (b) cluster dialogue topics, and (c) classify speaking mode, e.g. dictation versus question-answers.

Clinic Visit Dialogue2note Sentence Alignment Creating a corpus of aligned clinic visit conversation dialogue sentences with corresponding clinical note sentences is instrumental for training language generation systems. Early work in this domain includes that of (Finley et al., 2018), which uses an automated algorithm based on some heuristics, e.g. string matches, and merge conditions, to align dictation parts of clinical notes. In (Yim et al., 2020), we annotated manual alignments between dialogue sentences and clinical note sentences for the entire visit; however, the dataset was small and artificial (66 visits). Here we utilize this approach on real data and additionally provide an automatic sentence alignment baseline system. To our knowledge, this is the first work to propose an automated sentence alignment system for entire clinic visit dialogue and note pairs.

Clinical Language Generation from Conversation (Finley et al., 2018) produced dictation parts of a report, measuring performance both on gold standard transcripts and raw ASR output using statistical MT methods. In (Liu et al., 2019), the authors labeled a corpus of 101K simulated conversations and 490 nurse-patient dialogues with artificial short semi-structured summaries. They experimented with different LSTM sequence-to-sequence methods, various attention mechanisms, pointer generator mechanisms, and topic information additions. (Enarvi et al., 2020) performed similar work with sequence-to-sequence methods on a corpus of 800K orthopaedic ASR generated transcripts and notes; (Krishna et al., 2020) on a corpus of 6862 visits of transcripts annotated with clinical note summary sentences. Unlike most of previous works, our task generates clinical note sentences from labeled transcript snippets, which are at times overlapping and discontinuous. (Krishna et al., 2020)'s CLUSTER2SENT oracle system does use gold standard transcript "clusters", though different from our setup, outputs entire sections. While this strategy presupposes an upstream conversation topic segmentation system[1] as well as some extractive summarization, generation based on smaller text chunks can lead to more controllable and accurate natural language generation, critical characteristics in health applications.

4 Corpus Creation

Data The data set was constructed from clinical encounter visits from 500 visits and 13 providers. The data for each visit consisted of a visit audio and clinical note. For each visit audio, speaker roles (e.g. clinician patient) were segmented and labeled. Automatically generated speech to text for each audio was manually corrected by annotators. Table 3 gives the summary statistics of the extracted visit audio. For all specialties, the average number of turns and sentences for transcript was 175 ± 111 and 341 ± 214, for a total of 87725 turns and 170546 sentences. The number of sentences for clinical note was 47 ± 24, for a total of 23421 sentences. Table 4 shows the number of turns and sentences per different types of speakers.

We also combined our data with external data, the mock patient visit (MPV) dataset, from (Yim

[1] A system that divides conversations into segments according to topics

et al., 2020) to create a total of 566 visits.[2]

specialty	providers	visits	duration	speakers
ENT	1	68	10 ± 4	4 ± 1
HAND	1	43	10 ± 4	3 ± 1
ORTHO	1	27	11 ± 5	4 ± 1
PODIATRY	4	174	7 ± 4	3 ± 1
PRIMARY	6	188	17 ± 9	4 ± 1
TOTAL	13	500	12 ± 8	4 ± 1

Table 3: Source audio statistics

Annotations Each annotation is based on a clinical note sentence association with multiple transcript sentences. A note sentence can be associated with zero transcript sentences and an INFERRED-OUTSIDE label for default template values, e.g. "cv: normal". One may also be associated with sets of transcript sentences and a set tag, e.g. DICTATION or QA (described below). Finally, when multiple sets have anaphoric references, they may be tied together using a GROUP label. Given this hierarchy, the annotation related to a single note sentence can be represented as a tree as shown in Figure 1.

Set labels

COMMAND: Spoken by the clinician to the scribe to make a change to the clinical note structure, e.g. *"add skin care macro."*

DICTATION: Spoken by the clinician to the scribe where the output text is expected to be almost verbatim, though with understood changes in abbrevations, number expressions, and language formatting commands, e.g. *"return in four to five days period."*

STATEMENT2SCRIBE: Spoken by the clinician to the scribe where information is communicated informally, e.g. *"okay so put down heart and lungs were normal"*

STATEMENT: Statements spoken by any participant in a clinic visit in natural conversation, e.g. *"it lasted about a week."*

speaker	sentences	turns
clinician_primary	99421	42480
patient	56052	36059
other	15073	9186
TOTAL	170546	87725

Table 4: Speaker statistics

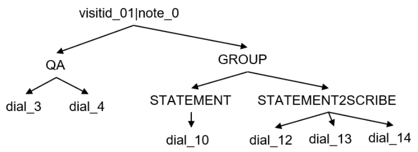

Figure 1: Annotation match tree

QA: Questions and answers spoken by any participant in a clinic visit in natural conversation, e.g. *"how long has the runny nose lasted? about a week."*

INFERRED-OUTSIDE: Clinical note sentences for which information comes from a known template's default value rather than the conversation, e.g.*"skin: intact."*

If after applying all possible associations and still there is information in the note sentence not available from the transcript, then an INCOMPLETE tag is added. A note sentence is left unmarked if no information can be found from the transcript. Table 2 shows label annotations with color coding for a full abbreviated transcript-note pair.

To measure interannotaor agreement, we calculated the triple, path, and span metrics introduced in (Yim et al., 2020), briefly described again here. The triple, path, and span metrics were defined based on instances constructed from the annotation tree representation. Specifically, for the triple metric, which measures unlabeled note to dialogue sentence match, instances are defined by note sentence id and transcript sentence id per visit, e.g. 'visitid_01|note_0|3'. The second metric, similar to the leaf-ancestor metric used in parsing, takes into account the full path from one note sentence to one dialogue sentence, e.g. 'visitid_01|note_0|GROUP|QA|3'. The span metric, similar to that of PARSEVAL, measures a node-level labeled span of dialogue sentences, e.g. for the top group node would be 'visitid_01|note_0|GROUP|[10,12,13,14]' (Sampson and Babarczy, 2003). When testing agreement, labels for each annotator are decomposed to these instance collections; true positive, false positive, and false negatives may be counted by the matches and mismatches between annotators. F1 score is calculated as usual. The different definitions allow both relaxed (triple) and stricter (path and span) agreement measurements.

[2] To normalize for annotation differences between the Mock Patient Visits (MPV) and our corpus, we removed INFERRED-DIALOGUE labels, reattached REPEATS to a higher node, and moved all GROUP labels to the highest node.

Labeling Process A group of 11 annotators were trained for various parts of the processing task. Audio transcription was performed using Elan (archive.mpi.nl/tla/elan) and dialogue2note annotation was performed using an in-house software. Annotators underwent training on sample files for which they received in-depth feedback. They additionally took a training quiz and self-reviewed errors. After training, their interannotator agreement was calculated based on 10 final files. Their average pairwise triple, path, and span F1 scores were 0.754, 0.549, and 0.645 respectively, a reasonable performance given the task difficulty.[3]

Annotation Statistics On average 58 ± 18 % of the clinical note was marked with an annotation. This suggests that around 40% of the note is structural, e.g. blank sentences or section headers, or from outside sources, e.g. injected labs, medication lists, etc. On average 13 ± 12 % of the transcript sentences were marked. This low number suggests that much of the information from transcripts consisted of repeats or were unused. Table 5 shows that most note sentences were associated with one set type, though still many were associated with multiple. Table 6 shows the frequency of note sentences and the unique label types associated with it. From the spread of percentages for each combination category, it is apparent that understanding the entire conversation context requires combining different types of cognitive listening skills. For each note sentence, the average range of transcript sentences associated with it in the train set was 11, with the 90th percentile at 17; however there were 10% of cases with ranges above 17, which occurred when explicit topic mentions appeared far away from detailed discussion. Crossing annotations occur when content from the note and transcript appeared comparatively out of order. For example, if note sentence 0 is matched with transcript sentence 3 and meanwhile note sentence 3 is matched with transcript sentence 0, these annotations "cross", rendering automatic alignment more challenging. To quantify this, we calculate the percentages of annotations which annotates cross one, three, or five other annotations[4] (Table 7). These high percentages reveal that the order of information in the transcript differ greatly from that of the note – thus alignments are said to be non-monotonic.

# label-types	freq	%
1	8712	37
2	2914	12
3	1021	4
4	311	1
5	20	-

Table 5: Label frequency per note sentence

label-combo	note sents	%sent	% cum
{INFERRED-OUTSIDE}	3731	16	16
{STATEMENT2SCRIBE}	2664	11	27
{STATEMENT}	977	4	31
{STATEMENT2SCRIBE,INCOMPLETE}	898	4	35
{DICTATION}	742	3	38
{STATEMENT,INCOMPLETE}	706	3	41
{QA}	465	2	43
{STATEMENT,GROUP}	452	2	45
{QA,STATEMENT,GROUP}	382	2	47

Table 6: Note sentence label combination statistics

The full amount of annotations from the dialogue2note labels may be used to create classifiers in many different types of tasks, e.g. dialogue relevance classification, topic segmentation, command identification, etc. However, in the remaining sections, we focus on two particular system applications : automatic dialogue2note sentence alignment and snippet summarization. For these baselines, the train and test sets were split using stratified random sampling using an 80-20 split. The training and test sets were composed of 400 and 100 of our visits; 53 and 13 for the MPV visits. 91 visits from training was reserved for development testing. As a simplification, the GROUP, INCOMPLETE, and COMMAND labels are ignored for these baselines.

crossing	percentages
cross1	33 ± 28
cross3	22 ± 27
cross5	14 ± 22

Table 7: Crossing annotation statistics

5 Sentence Alignment Baselines

We define the dialogue2note sentence alignment baseline task as the classification of 1-to-1 dialogue sentence and clinical note sentence pairs with set labels. Thus, the candidate space includes all combinations of clinical note sentences paired with all dialogue possible sentences in a visit; only those annotated with labeled associations are considered positive. This is a subset of the full annotation tasks that require 1-to-many multi-label classifications with hierarchical GROUP set labels. However, this

[3] These agreement values are consistent with the comparable task of simplification corpus creation, previously measured to be 0.68 kappa (Hwang et al., 2015).

[4] DICTATION, STATEMENT2SCRIBE, COMMAND labels aren't counted to focus on conversational dialogue

feature	description
match-note	Dot product of note and transcript vector divided by the magnitude of the note vector.
match-transcript	Dot product of note and transcript vector divided by the magnitude of the transcript vector.
cui-pair	UMLS concept pair, as extracted by MetaMap (Aronson and Lang, 2010), where the first concept unique identifier (cui) is from the clinical note and the second cui is from the transcript sentence. The **top_p** parameter determines which most significant cui-pair features to keep, using chi-square analysis.
prev-sent-quest	1 if the previous sentence has one of sentence has a question feature, e.g. interrogative words such as who, what etc, 0 otherwise.
jaccard-sim	If set to **local**, then defaults to jaccard similarity of the note-transcript sentence pair. If set to **regional** and similarity passes the **sim-thresh** threshold, instead, the maximum jaccard similarity from candidate regional local matches is returned. These candidate regional matches are created by by heuristically finding the closest length matches by incorporating previous and next sentences.

Table 8: Feature description for non-standard features

setup is consistent with the comparable simplification dataset creation task. We report the alignment evaluation based on pairwise F1 score. The number of positive pairwise instances in train, dev, and test sets are 19721, 4770, and 5796; including all possible negative instances 6370787, 1303972, 1706901.

Bitext Corpus Creation Related Work The topic of bitext corpus creation is often used in the context of creating resources for statistical machine translation or as a means to create cross lingual linguistic resources (Koehn, 2005; Tiedemann, 2011); it is also used to describe simplification dataset creation (Barzilay and Elhadad, 2003; Hwang et al., 2015; Štajner et al., 2017). While highly parallel bitext can be aligned using sentence length methods, much like other comparable corpora alignment strategies, multi-form comparable corpora cannot rely on monotonic ordering or correlated bitext sentence length; moreover the different text forms presents additional constraints on exact narrative structure. Like in previous work, we build our baselines for dialogue2note sentence alignment by using similarity features with some adjustment to incorporate similarity over multiple sentences.

System Description Candidate classification instances for every note sentence and transcript sentences pair were created and classified into one of the previously described set labels. For each clinical note, an additional classification instance was created for a match with an empty transcript line. (This occurs with a INFERRED-OUTSIDE label). A single tag was assigned to each classification instance according to annotated labels. If multiple tags existed per sentence pair, we took the first label in the following order: STATEMENT, STATEMENT2SCRIBE, QA, DICTATION.

Sentences were tokenized, changed to lemma form using Spacy English model (spacy.io), and vectorized according to a bag of words model. Stop words and punctuation were removed. To balance the uneven data distribution, the number of negative class instances were sampled randomly according to configurable parameter, neg_samp. We experimented with three baseline pairwise classification systems:

simple-threshold : A rule-based system that categorizes everything over threshold1 to DICTATION anything between threshold1 and threshold2 to STATEMENT2SCRIBE. These were the two labels in the train set with the highest pairwise similarities; other labels had comparable similarities.
system1 : A simple feature-based system using a decision tree classifier (scikit-learn.org). Its features included speaker category, cosine similarity, length of the note and transcript sentence vectors, and the note sentence vector. In order to take into account the match over the length of either the note or the transcript, we included a match-note and match-transcript feature described in Table 8.
system2 : A feature-based system like system1 with additional features, the transcript vector, a previous-question feature, a cui-pair feature, and a jaccard similarity feature described in Table 8. To avoid erroneous matches to answer sentences, in this system, common answers (e.g. "no") were removed from the train set.

Results After tuning, we found optimal performances for the threshold systems at threshold1=0.9 and threshold2=0.6. For system1 and system2, optimized parameters were at neg_samp=50, jaccard-sim=regional, sim-thresh=0.3, top_p=20, for a decision tree classifier. Table 9 shows the F1 results per each label. With the simple threshold system, we can see the DICTATION pairs already achieve a

label	thresh	sys1	sys2
DICTATION	0.36	0.39	0.43
STATEMENT2SCRIBE	0.20	0.36	0.36
STATEMENT	0.00	0.12	0.13
QA	0.00	0.19	0.20
INFERRED-OUTSIDE	0.00	0.59	0.66
UNMARKED	0.998	0.998	0.998

Table 9: Pairwise F1 by label

similarity	composition	thresh	sys1	sys2
0-20	0.66	0.00	0.22	0.26
20-40	0.20	0.08	0.39	0.39
40-70	0.09	0.45	0.64	0.69
70-100	0.05	0.91	0.94	0.93

Table 10: Pairwise F1 by jaccard similarity (composition is the percent of annotations within the range)

performance near that of the more complex systems. Using a simple feature based system, we see F1 measures between 0.188 and 0.390 for everything but INFERRED-OUTSIDE and UNMARKED. As expected, given the high amounts of UNMARKED, it has the highest performance. Adding additional features and curating training examples gave a minor boost across different labels as shown in the system1 and system2 differences. Analyzing the results across pairs based on similarity ranges, we see that the higher similarity pairs have higher performance, likely because the similarity features can be more reliable at those ranges (Table 10). Table 11 shows the results of system2 per label. Such results are comparable to simplification dataset creation systems with 0.33 F1 at 0-40% similarity, 0.79 F1 at 40-70%, 0.95 F1 at 70-100% (Barzilay and Elhadad, 2003).

label	gold-freq	P	R	F1
DICTATION	257	0.53	0.35	0.43
STATEMENT2SCRIBE	1248	0.32	0.43	0.36
STATEMENT	2140	0.23	0.09	0.13
QA	1239	0.25	0.16	0.20
INFERRED-OUTSIDE	912	0.72	0.61	0.66
UNMARKED	1701105	0.998	0.998	0.999

Table 11: Sys2 performance by label

Studying confusions between classes in system2, we found that overwhelmingly most errors were due to assigning unmarked passages to another label. This may be due to the simple representation of features, where certain content note or transcript bag of word features may have higher weights against similarity features. There are also cases where legitimately, the dialogue will mention what is discussed in the clinical note but is not marked in the gold standard (e.g. the same topic may be referred to multiple times but we only annotate the best instance). To a smaller extent, there were confusions among related positive class labels. Confusions between DICTATION and STATEMENT2SCRIBE occurred for high similarity sentences. Confusions between STATEMENT2SCRIBE and STATEMENT arose for cases in which dialogue may be perceived to be spoken either to a scribe or a patient, e.g. “looks normal”. Confusions between STATEMENT and QA transpired because we allowed the QA label to encompass both open-ended questions, e.g. “How are you? I have been having a headache for 2 weeks” as well as very focused categorical questions, e.g. “Did you take nasal spray? No.”; thus answers to open-ended questions can be easily confused with STATEMENTs.

In the current system, classifications for each note-dialogue sentence pair are labeled independently. We can improve the system by framing the required matches for each clinical note sentence as a sequence labeling problem. More semantic normalization features and surrounding sentence features would benefit the classification. Finally, in the future we can try more complex sentence vector representations.

6 Snippet Summarization Baselines

We define the snippet summarization baseline task where given the gold standard dialogue snippet text, a corresponding clinical note sentence is generated. The number of instances of aligned sets for train, dev, and test was 7129, 1851, and 2085 respectively. The average number of input and output tokens was 24 and 13 respectively.

Monolingual Text-to-Text Language Generation Related Work Monolingual monologue text-to-text language generation tasks include summarization (See et al., 2017), simplification (Štajner et al., 2017), and paraphrasing (Ma et al., 2018). The exact manner of transformation between the input and output text depends on comparative lengths, task-specific constraints, and level of abstraction.

In the area of conversational modeling, e.g. chatbots, the task is to produce appropriate dialogue responses given a prompt. In one simple classic setup, the response generation can be modeled as an information retrieval problem (Jurafsky and Martin, 2009; Ji et al., 2014). In such systems, the prompt query is processed and compared to those saved

	BLEU				**R-1**				**R-2**				**R-L**			
section	**ret**	**vanilla**	**pg**	**pg-mt**	**ret**	**vanilla**	**pg**	**pg-mt**	**ret**	**vanilla**	**pg**	**pg-mt**	**ret**	**vanilla**	**pg**	**pg-mt**
AP	0.26	0.19	**0.38**	**0.38**	0.20	0.14	**0.33**	0.31	0.09	0.03	0.17	**0.18**	0.19	0.13	**0.31**	0.30
CC	0.22	0.22	**0.30**	**0.30**	0.16	0.15	**0.26**	0.24	0.05	0.05	**0.11**	0.10	0.15	0.14	**0.25**	0.22
HPI	0.26	0.19	**0.36**	**0.36**	0.18	0.16	**0.32**	0.29	0.07	0.04	**0.14**	**0.14**	0.17	0.15	**0.30**	0.27
IM	0.29	0.15	0.61	**0.73**	0.42	0.17	0.65	**0.75**	0.31	0.02	0.52	**0.61**	0.41	0.16	0.63	**0.73**
PE	0.35	0.19	**0.44**	**0.44**	0.29	0.17	0.39	**0.40**	0.17	0.04	**0.24**	0.23	0.28	0.16	0.38	**0.39**
ROS	0.15	0.12	0.21	**0.22**	0.13	0.11	**0.28**	0.24	0.03	0.01	0.04	**0.06**	0.11	0.10	**0.27**	0.22
ALL	0.27	0.19	**0.38**	**0.38**	0.21	0.15	**0.33**	0.32	0.09	0.04	**0.17**	**0.17**	0.19	0.14	**0.32**	0.30

Table 12: BLEU, ROUGE-1, ROUGE-2, and ROUGE-L performance by sections

in training data. The system produces the saved response to the prompt most similar to that of the query. Although our task is not to respond a user, we may utilize the same type of system. Specifically, we can instead model the note sentence as the retrieval response to a dialogue input prompt.

Our problem most closely resembles meeting conversation summarization, in which the source data is a meeting conversation (dialogue) and the target data is a meeting summary (monologue) (Carenini et al., 2011). Method pipelines include multiple classifiers such as topic segmentation, action item identification, as well as some language generation module. There is also work with end-to-end pipelines that perform extractive and abstractive neural generation (Zhu et al., 2020; Mehdad et al., 2013). Unlike a typical summarization task, our source data is of a more comparable length, making the task more tractable. For our baselines, in addition to a simple retrieval based system, we experimented with a classic sequence-to-sequence model with and without a pointer-generator.

Note Section Identification Clinical notes are typically organized into different sections demarked by section headers as shown in Table 2 note lines 0, 2, 26, and 62. In order to report language generation performances grouped by sections and also to experiment with joint section prediction, we automatically labeled note sentences to one of six note sections using a rule-based algorithm. These categories included: History of Present Illness (HPI), Assessment and Plan (AP), Physical Exam (PE), Chief Complaint (CC), Review of Systems (ROS), and Imaging (IM). Sections headers were identified using regular expressions created by studying the train set. Subsequently, note sentences were labeled based on their corresponding section header. We modeled section prediction for two of our baseline systems : **ret**, **pg-mt**.

System Descriptions Below we describe our baseline systems. We trained and tested our seq-to-seq models using the LeafNATS codebase (Shi et al., 2019).

retrieval-based generator (**ret**) : Note sentence suggestion generation are modeled as a retrieval task. Paired transcript snippets and note lines (with associated section) are cached. For new transcript snippets, the note sentence corresponding to the highest cosine similarity dialogue snippet in training data is returned.

seq2seq baselines : We evaluate the performances of three sequence-to-sequence baselines with an RNN sequence encoder. The base system (**vanilla**) is a simple sequence-to-sequence system with attention. We also evaluate an option to add a pointer-generator network (**pg**). Finally, to model a pointer-generator system that outputs a summary as well as a section designation, we evaluated a final option that treats the two outputs as a multitask system (**pg-mt**).[5] Experiments were run on an EC2 p2.xlarge instance with an NVIDIA K80 GPU, taking ∼150 minutes each.

Results Table 12 shows the BLEU, ROUGE-1 (R-1), ROUGE-2 (R-2), AND ROUGE-L (R-L) performances across different note sections. As shown, typically the two pointer-generator systems outperform the retrieval based and vanilla baselines. This difference may be due to the ability for the pointer-generator system to copy-and-paste items from the original input.

Comparatively, (Krishna et al., 2020)'s best CLUSTER2SENT oracle scores yielded R-1, R-2, and R-L performances of 66.5, 39.01, and 52.46, respectfully, from 6862 visits. In our low resource scenario of 566 visits, we achieved 50%, 43%, and 61% of their R-1, R-2, and R-L scores at 12% of the data. This suggests given more training data our

[5] Final experimental hyperparameters were set at, RNN=LSTM, batch_size=50, emb_dim=128, src_hidden_dim=256, trg_hidden_dim=256, src_seq_lens=400, trg_seq_lens=100, attn_method=luong_concat, repetition=vanilla, share_emb_weight=False.

system may similarly reach state-of-the-art levels.

Table 13 shows the accuracy of the **ret** and **pg-mt** systems for note section prediction. Although on the whole, **pg-mt** performs better than the **ret** system, for low frequency categories this is not the case. This phenomenon most likely occurs because **pg-mt** favors higher frequency labels, which is consistent with its training objective. **ret**, which classifies note section through the intermediate comparisons of input sequence similarities, is less likely to be directly skewed by class imbalances.

section	freq			acc	
	train	validation	test	ret	pg-mt
AP	1935	534	655	0.41	0.54
CC	306	71	113	0.12	0.00
HPI	3708	949	956	0.65	0.85
IM	85	7	0	0.38	0.00
PE	992	274	319	0.59	0.58
ROS	103	16	21	0.05	0.00
ALL	7129	1851	2085	0.53	0.65

Table 13: Section frequency and accuracy

Human Evaluation We sampled 10 random test snippets from each of the six section categories for evaluation (total 60 snippets). An annotator with a medical degree was asked to rank the four systems relative to each other, where 1 is the best. Additionally each system was evaluated independently with a score from 1-5 (5=best) for the categories relevancy, factual accuracy, writing-style, completeness, and overall. Table 14 shows the average scores for the different baseline systems. The vanilla seq2seq system consistently performed the worst, while the pointer-generator systems consistently performed better.

	ret	vanilla	pg	pg-mt
completeness	2.5	1.2	3.1	2.9
factual-accuracy	2.4	1.3	3.2	2.9
relevancy	2.9	1.5	3.7	3.5
writing-style	3.2	1.8	3.3	3.3
overall	2.4	1.2	3.1	2.9
rank(1=best)	2.7	3.4	1.8	2.1

Table 14: Average human evaluation ratings

While our sentence generation baselines showed modest performances, this is consistent with low resource language generation scenarios and may be ameliorated with additional training data. To improve our system, in the future, we will apply methods from low-resource machine translation techniques, utilizing unpaired sources of medical dialogue and clinic note corpora. Furthermore, we can experiment with other sequence-to-sequence approaches, e.g. transformers, for better summary generation. Joint section prediction generation may be extended to model hierarchical sections by adjusting targets to include subsections.

7 Conclusions

In this work, we provided baselines for two tasks that work towards natural language generation of note sentences from medical visit conversation. An automated dialogue2note sentence alignment system can be used to create realistic training data so immensely critical for modern systems. Meanwhile, if given properly extracted transcript snippets, dialogue2note snippet summarization could provide a valuable building block for an overall language generation system.

In future work, additional metadata information, (e.g. set labels, speaker, specialties) may be incorporated into the network architecture. Although we only explore two systems here, other models such as topic segmentation, extractive summarization, note sentence ordering, and dialogue command classification, can be trained from this annotated dataset alone. These labels may alternatively be used for additional multitask classification objectives in a full sequence-to-sequence model.

Extension of this labeled dataset may yield further interesting gains. For example, textual entailment labels between paired snippets would allow progress towards understanding and generating semantic variations and detail. Event annotation, which structures text, if performed on paired snippets, would provide training examples for data-to-text or text-to-data generation.

Together or apart, such systems would enable automation of clinical note generation whether as a full end-to-end solution or as piecemeal suggestions in a human-augmented solution. Ultimately this technology may be utilized to deburden clinicians, allowing them to focus back on patient care.

Ethical Considerations

All annotators, hired in-house, underwent HIPAA data and security training. Data was stored in dedicated HIPAA compliant compute resources. Data collection and persistence was consistent with terms of use and customer expectations. All content examples in this paper are fictitious.

References

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

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

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

Giuseppe Carenini, Gabriel Murray, and Raymond Ng. 2011. Methods for mining and summarizing text conversations. *Synthesis Lectures on Data Management*, 3(3):1–130. Publisher: Morgan & Claypool Publishers.

Miroslav Cerny. 2007. On the function of speech acts in doctor-patient communication. *Linguistica*.

Kriss A Drass. 1982. Negotiation and the structure of discourse in medical consultation. *Sociology of health& illness*.

Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. Extracting symptoms and their status from clinical conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 915–925. Association for Computational Linguistics.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30. Association for Computational Linguistics.

Gregory Finley, Wael Salloum, Najmeh Sadoughi, Erik Edwards, Amanda Robinson, Nico Axtmann, Michael Brenndoerfer, Mark Miller, and David Suendermann-Oeft. 2018. 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. Association for Computational Linguistics.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard wikipedia to simple wikipedia. In *HLT-NAACL*.

Serena Jeblee, Faiza Khan Khattak, Noah Crampton, Muhammad Mamdani, and Frank Rudzicz. 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.

Zongcheng Ji, Z. Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *ArXiv*, abs/1408.6988.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation.

K. Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary Chase Lipton. 2020. Generating soap notes from doctor-patient conversations. *ArXiv*, abs/2005.01795.

Ronilda C. Lacson, Regina Barzilay, and William J. Long. 2006. Automatic analysis of medical dialogue in the home hemodialysis domain: Structure induction and summarization. *Journal of Biomedical Informatics*, 39(5):541–555.

Zhengyuan Liu, A. Ng, Sheldon Lee Shao Guang, AiTi Aw, and Nancy F. Chen. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821.

Shuming Ma, Xu Sun, Wei Li, Sujian Li, Wenjie Li, and Xuancheng Ren. 2018. Query and output: Generating words by querying distributed word representations for paraphrase generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 196–206, New Orleans, Louisiana. Association for Computational Linguistics.

Yashar Mehdad, G. Carenini, F. Tompa, and R. Ng. 2013. Abstractive meeting summarization with entailment and fusion. In *ENLG*.

Juan C. Quiroz, Liliana Laranjo, Ahmet Baki Kocaballi, Shlomo Berkovsky, Dana Rezazadegan, and Enrico Coiera. 2019. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digital Medicine*, 2.

Ginette Raimbault, Olga Cachin, Jean Marie Limal, Caroline Eliacheff, and Raphael Rappaport. 1975. Aspects of communication between patients and doctors: an analysis of the discourse in medical interviews. *Pediatrics*.

Alvin Rajkomar, Anjuli Kannan, Kai Chen, Laura Vardoulakis, Katherine Chou, Claire Cui, and Jeffrey Dean. 2019. Automatically charting symptoms from patient-physician conversations using machine learning. *JAMA Internal Medicine*, 179(6):836–838.

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

Benjamin J Schloss and Sandeep Konam. 2020. Towards an automated SOAP note: Classifying utterances from medical conversations. In *Proceedings of Machine Learning Research, Machine Learning for Healthcare (MLHC)*.

A. See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Tian Shi, Ping Wang, and Chandan K Reddy. 2019. Leafnats: An open-source toolkit and live demo system for neural abstractive text summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 66–71.

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

Brian D. Tran, Yunan Chen, Songzi Liu, and Kai Zheng. How does medical scribes' work inform development of speech-based clinical documentation technologies? a systematic review. 27(5):808–817.

Nan Wang, Yan Song, and Fei Xia. 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. European Languages Resources Association (ELRA).

Nan Wang, Yan Song, and Fei Xia. 2020. Studying challenges in medical conversation with structured annotation. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 12–21. Association for Computational Linguistics.

Wen-wai Yim, Meliha Yetisgen, Jenny Huang, and Micah Grossman. 2020. Alignment annotation for clinic visit dialogue to clinical note sentence language generation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 413–421. European Language Resources Association.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. End-to-end abstractive summarization for meetings. *ArXiv*, abs/2004.02016.

Sanja Štajner, Marc Franco-Salvador, Simone Paolo Ponzetto, Paolo Rosso, and Heiner Stuckenschmidt. 2017. Sentence alignment methods for improving text simplification systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 97–102. Association for Computational Linguistics.