Extracting relevant information from physician-patient dialogues for automated clinical note taking

Serena Jeblee[1,2], Faiza Khan Khattak[1,2,3], Noah Crampton[3],
Muhammad Mamdani[3], Frank Rudzicz[1,2,3,4]
[1]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[2]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada
[3]Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Ontario, Canada
[4]Surgical Safety Technologies, Toronto, Ontario, Canada
sjeblee@cs.toronto.edu, faizakk@cs.toronto.edu,
cramptonn@unityhealth.to, muhammadm@unityhealth.to, frank@spoclab.com

Abstract

We present a system for automatically extracting pertinent medical information from dialogues between clinicians and patients. The system parses each dialogue and extracts entities such as medications and symptoms, using context to predict which entities are relevant. We also classify the primary diagnosis for each conversation. In addition, we extract topic information and identify relevant utterances. This serves as a baseline for a system that extracts information from dialogues and automatically generates a patient note, which can be reviewed and edited by the clinician.

1 Introduction

In recent years, electronic medical record (EMR) data have become central to clinical care. However, entering data into EMRs is currently slow and error-prone, and clinicians can spend up to 50% of their time on data entry (Sinsky et al., 2016). In addition, this results in inconsistent and widely variable clinical documentation, which present challenges to machine learning models.

Most existing work on information extraction from clinical conversations does not differentiate between entities that are relevant to the patient (such as experienced symptoms and current medications), and entities that are not relevant (such as medications that the patient says were taken by someone else).

In this work, we extract clinically relevant information from the transcript of a conversation between a physician and patient (and sometimes a caregiver), and use that information to automatically generate a clinical note, which can then be edited by the physician. This automated note-taking will save clinicians valuable time and allow them to focus on interacting with their patients rather than the EMR interface. We focus on linguistic context and time information to determine which parts of the conversation are medically relevant, in order to increase the accuracy of the generated patient note. In addition, the automatically generated notes can provide cleaner and more consistent data for downstream machine learning applications, such as automated coding and clinical decision support.

Figure 1 shows a synthetic example of the kind of medical conversation where context and time information are important.

DR: Are you currently taking [Adderall]$_{\text{Med.}}$?
PT: No, but I took it [a few years ago]$_{\text{TIMEX3}}$.
DR: And when was that?
PT: Um, around [2015 to 2016]$_{\text{TIMEX3}}$.
DR: And did you ever take [Ritalin]$_{\text{Med.}}$?
PT: I dont think so.

Typical output: *Adderall, Ritalin.*
Expected output:
Medications: Adderall (2015-2016), no Ritalin

Figure 1: Synthetic conversation example.

2 Related Work

Previous studies have shown that current EMR data are difficult to use in automated systems because of variable data quality (Weiskopf and Weng, 2013; Thiru et al., 2003; Roth et al., 2003). Weiskopf and Weng (2013) showed that EMR data is frequently incomplete, and is often not evaluated for quality. In addition, the variance in documentation style, abbreviations, acryonyms, etc. make it difficult for algorithms to interpret the text.

Some recent work on machine learning methods for EMR data includes predicting mortality

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 65–74
Hong Kong, November 3, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/D19-62

and discharge diagnoses (Rajkomar et al., 2018), predicting unplanned hospital readmissions for 5k patients by encoding EMR data with a convolutional neural network (Nguyen et al., 2018), and predicting diagnosis codes along with text explanations (Mullenbach et al., 2018).

Although there is some existing work on generating text from structured data (Dou et al., 2018; Lebret et al., 2016), very little work has been done in the clinical domain. Liu (2018) generated patient note templates with a language model, which was able to approximate the organization of the note, but no new information from the patient encounter was used.

Du et al. (2019) introduced a system for extracting symptoms and their status (experienced or not) from clinical conversations using a multi-task learning model trained on 3,000 annotated conversations. However, their model was trained on a limited set of 186 symptoms and did not address other medically relevant entities.

A latent Dirichlet allocation (LDA) model (Blei et al., 2003) is a topic modeling technique and has been applied to clinical text to extract underlying useful information. For example, Bhattacharya et al. (2017) applied LDA on structured EMR data such as age, gender, and lab results, showing that the relevance of topics obtained for each medical diagnosis aligns with the co-occurring conditions. Chan et al. (2013) applied topic modeling on EMR data including clinical notes and provided an empirical analysis of data for correlating disease topics with genetic mutations.

3 Dataset

Primary diagnosis	Dyads
ADHD	100
Depression	100
COPD	101
Influenza	100
Osteoporosis	87
Type II diabetes	86
Other	226

Table 1: Data distribution (ADHD: Attention Deficit Hyperactivity Disorder; COPD: Chronic Obstructive Pulmonary Disorder)

For training and testing our models, we use a dataset of 800 patient-clinician dialogues (dyads) purchased from Verilogue Inc.[1], which includes demographic information about the patient as well as the primary diagnosis. The data consist of audio files and human-generated transcripts with speaker labels. Table 1 shows the distribution of diagnoses in the dataset.

Since these data are proprietary, we also use a few transcripts of staged clinical interviews from YouTube as examples. [2]

4 Annotation

First, the conversation transcripts are automatically annotated for time phrases using HeidelTime, a freely available rule-based time phrase tagger (Strötgen and Gertz, 2010), as well as a limited set of common medical terms.

Two physicians then conduct manual annotation by correcting the automatic annotations and making any necessary additions, using a custom-developed annotation interface. The following types of entities are annotated: anatomical locations, diagnoses, symptoms, medications, reasons for visit, referrals, investigations/therapies, and time phrases. A total of 476 conversations are annotated by a unique physician, and inter-annotator agreement is calculated using DKPro Statistics[3] on 30 conversations which were annotated by both physicians. The agreement across all entity types is 0.53 Krippendorffs alpha (Krippendorff, 2004) and 0.80 F_1 (partial match).

We developed a custom annotation interface for labeling entities and their attributes in the transcripts, shown in Figure 3. The software includes the ability to add new annotation types and attributes, edit and delete previous annotations, and view the entire conversation for context.

[1] http://www.verilogue.com

[2] YouTube videos of simulated patient encounters were sourced by searching for the following terms: "medical history", "patient interview", and "clinical assessment". Our clinician team member watched potential videos in the search list and selected only the ones that met the following criteria: 1) clinician asking a patient questions in simulated clinical scenarios; 2) a subjective perception of adequate fidelity to real clinical encounters. The audio for these dialogues were transcribed by a professional transcriptionist. Examples used in this paper:
1: https://www.youtube.com/watch?v=O2qYU8n4VsA, 2: https://www.youtube.com/watch?v=CUSxC-XHT2A, 3: https://www.youtube.com/watch?v=5_jIcAk1XeA

[3] https://dkpro.github.io/dkpro-statistics

DR: *It's a shame how good the Blue Jays were a couple of seasons ago compared to now.* **PT:** *Yeah, I'm still not sure we should have got rid of Alex Anthopoulos.* **DR:** *Yeah, that was the turning point, eh? Anyways, you're here to review your [diabetes]*$_{\text{Diagnosis}}$ *right?* **PT:** *That's right.* **DR:** *How's the [numbness in your toes]*$_{\text{Sign/Symptom}}$*/[toes]*$_{\text{Anatomical Location}}$ *?* **PT:** *The same. I'm used to it by now, and I'm grateful it's not getting worse.* **DR:** *Okay, that's good. Let's keep you on the [same dose of Metformin]*$_{\text{Medication}}$ *[for now]*$_{\text{TIMEX3}}$ *then we'll check your [a1c]*$_{\text{Investigation/Therapy}}$ *again [in three months]*$_{\text{TIMEX3}}$ *, and then I'll [see you back here after that]*$_{\text{Disposition plan}}$. **PT:** *That makes sense to me.*	**DR:** *It's a shame how good the [Blue]*$_{\text{Medication}}$ *Jays were a couple of seasons ago compared to [now]*$_{\text{TIMEX3}}$. **PT:** *Yeah, I'm still not sure we should have got rid of Alex Anthopoulos.* **DR:** *Yeah, that was the turning point, eh? Anyways, you're here to review your [diabetes]*$_{\text{Diagnosis}}$ *right?* **PT:** *That's right.* **DR:** *How's the numbness in your [toes]*$_{\text{Anatomical Location}}$ *?* **PT:** *The same. I'm used to it by [now]*$_{\text{TIMEX3}}$ *,and I'm grateful it's not getting worse.* **DR:** *Okay, that's good. Let's keep you on the same [dose]*$_{\text{Medication}}$ *of [Metformin]*$_{\text{Medication}}$ *for [now]*$_{\text{TIMEX3}}$ *then we'll check your a1c again in [three months]*$_{\text{TIMEX3}}$ *, and then I'll see you back here after that.* **PT:** *That makes sense to me.*

Figure 2: Example dialogue: (Left) Human annotation, (Right) automatic annotation. In both tables, highlight indicates the annotated entities; darker highlights indicate overlap between human and automatic annotations. Subscripts indicate the entity type.

5 Methods and experiments

The automated pipeline currently includes the following components: preprocessing, utterance type classification (questions, answers, statements, etc.), entity extraction (medications, symptoms, diagnoses, etc.), attribute classification (modality and pertinence), primary diagnosis classification, SOAP classification, and note generation. In this section we discuss each component in detail, including methods and results. See Figure 4 for a diagram of the system components.

5.1 Preprocessing and data splitting

Before passing the data to our models, the text of the transcripts is lowercased, and punctuation is separated from words using WordPunctTokenizer from NLTK (Steven Bird and Loper, 2009). For the utterance type and attribute classification tasks, each word in an utterance is represented as a word embedding. In this work, we use publicly available ELMo embeddings (Peters et al., 2018) trained on PubMed abstracts, as well as word2vec embeddings trained on PubMed[4].

Of the 476 annotated conversations, we randomly select 50 to use as a test set for entity extraction and attribute classification.

5.2 Utterance type classification

In order to understand the conversational context, it may be useful to know whether an utterance is a question or answer. To this end, we classify each utterance as one of the following types: question, statement, positive answer, negative answer, backchannel (such as ‘uh-huh’ or ‘yeah’) or excluded (incomplete or vague utterance).

The utterance type classification model is a two-layer bidirectional gated recurrent unit (GRU) neural network (Cho et al., 2014), implemented in PyTorch, with the architecture shown in Figure 5. We augment the training data with two external, publicly available datasets: the Switchboard corpus (Calhoun et al., 2010), and the AMI corpus[5]. We map the utterance labels from the AMI and Switchboard corpora to our set of labels, and add these data to our training set.

We evaluate the utterance type classifier on a set of 20 conversations, annotated independently by 2 annotators with inter-annotator agreement of 0.77 (Cohen’s kappa).

Table 2 shows the classification results by utterance type. As the most frequent type, statements are the easiest for the model to identify. The low performance of infrequent classes indicates that we could potentially improve performance by using an oversampling or regularization method.

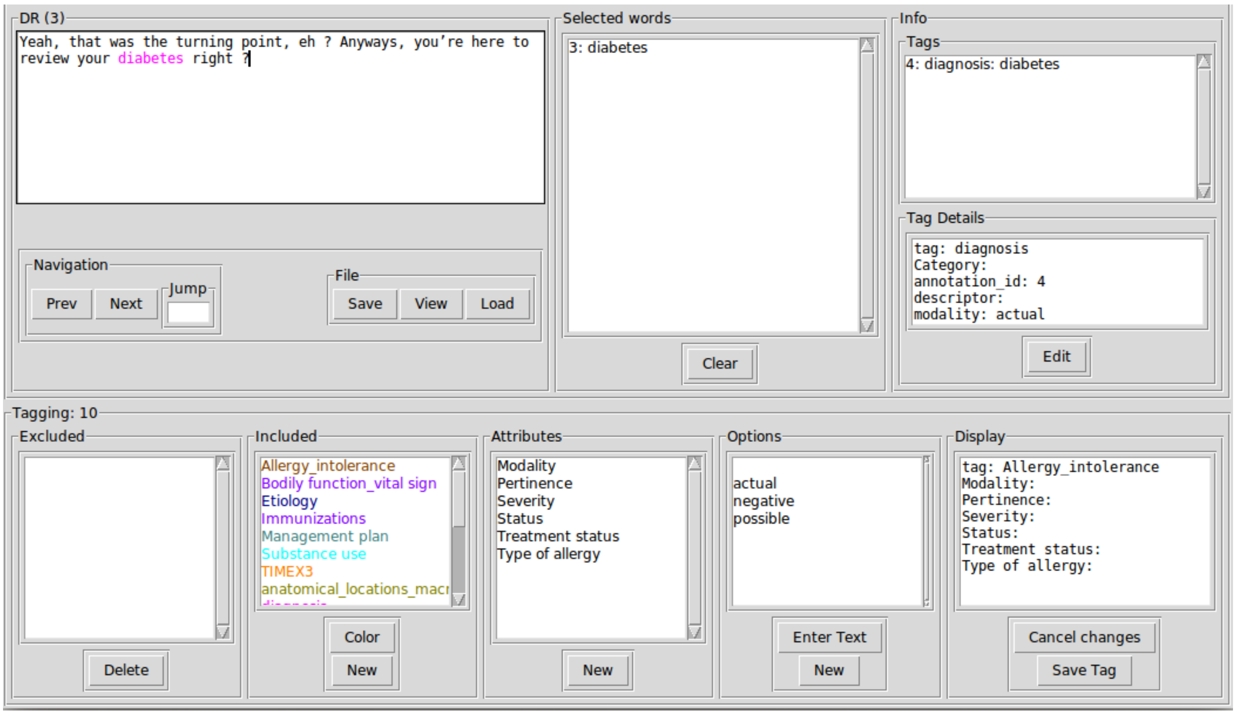
5.3 Entity extraction

5.3.1 Time phrase extraction

In order to determine clinical relevance, it is important to know the time and duration of events in the patient history. We use HeidelTime to identify

[4] http://evexdb.org/pmresources/vec-space-models/

[5] http://groups.inf.ed.ac.uk/ami/corpus/



Yeah, that was the turning point, eh ? Anyways, you’re here to review your diabetes right ?

3: diabetes

4: diagnosis: diabetes

```
tag: diagnosis
Category:
annotation_id: 4
descriptor:
modality: actual
```

Allowed_intolerance
Bodily function_vital sign
Etiology
Immunizations
Management plan
Substance use
TIMEX3
anatomical_locations_macr

Modality
Pertinence
Severity
Status
Treatment status
Type of allergy

actual
negative
possible

```
tag: Allergy_intolerance
Modality:
Pertinence:
Severity:
Status:
Treatment status:
Type of allergy:
```

Figure 3: Annotation interface

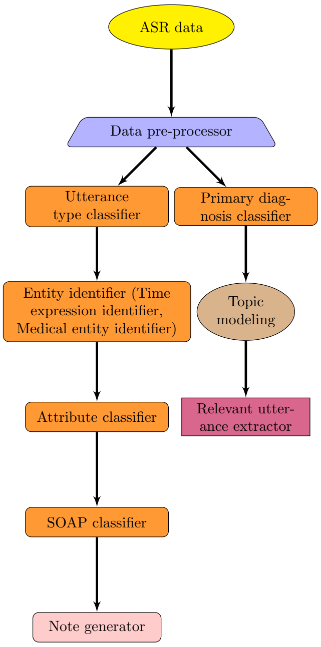

ASR data
Data pre-processor
Utterance type classifier
Primary diagnosis classifier
Entity identifier (Time expression identifier, Medical entity identifier)
Topic modeling
Attribute classifier
Relevant utterance extractor
SOAP classifier
Note generator

Figure 4: System components and data flow

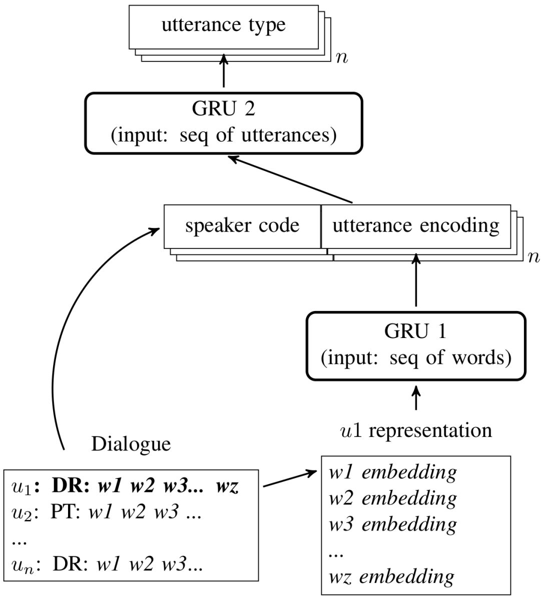


Figure 5: Utterance type classification model

time phrases in the transcripts, including times, dates, durations, frequencies, and quantities.

5.3.2 Clinical entity extraction

In addition to time phrases, we identify the following clinical concept types: anatomical locations, signs and symptoms, diagnoses, medications, referrals, investigations and therapies, and reasons for visit. To identify these entities, we search the transcript text for entities from a variety of medical lexicons, including the BioPortal Symptom lexicon [6], SNOMED-CT [7], the Consumer Health Vocabulary (CHV) [8], and RxNorm (a database of

[6] https://bioportal.bioontology.org/ontologies

[7] http://www.snomed.org/

[8] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/

Type	Instances	P	R	F_1
Question	539	0.72	0.49	0.59
Statement	2,347	0.82	0.83	0.82
AnswerPositive	195	0.36	0.41	0.38
AnswerNegative	82	0.74	0.34	0.47
Backchannel	494	0.56	0.76	0.64
Excluded	131	0.20	0.16	0.18
Average	3,788	0.72	0.72	0.71

Table 2: Utterance type classification results

normalized medication names) [9].

Entity identification is currently limited to the terms present in our reference lists, which are large but cannot cover all possible expressions of relevant entities. There may be many valid variations of these entities that we hope to be able to identify in the future, potentially using a more sophisticated tagging method such as named entity recognition (NER).

Type	Instances	P	R	F_1
Anatomical locations	328	0.79	0.45	0.57
Diagnosis	346	0.88	0.62	0.72
Investigation or therapy	239	0.42	0.24	0.31
Medication	579	0.55	0.79	0.65
Referral	61	0.11	0.11	0.11
Sign/symptom	650	0.82	0.38	0.52
Time expression	1286	0.98	0.64	0.77
Average	3489	0.80	0.56	0.64

Table 3: Entity extraction results

5.3.3 Attribute classification

After extracting relevant entities, we classify them according to several attributes, including modality (i.e., whether the events were actually experienced or not) and pertinence (i.e., to which disease the entities are relevant, if any). For example, a patient might mention a medication that they have not actually taken, so we would not want to record that medication as part of the patient's history. In these cases, the context of the conversation, as well as time information, is crucial to recording the patient's information accurately.

Type	Instances	P	R	F_1
Actual	504	0.87	0.80	0.83
Negative	144	0.63	0.64	0.64
Possible	5	0.09	0.40	0.14
None	91	0.59	0.71	0.65
Average	744	0.78	0.76	0.77

Table 4: Modality classification results

Type	Instances	P	R	F_1
ADHD	126	0.54	0.41	0.28
COPD	22	0.20	0.45	0.28
Depression	32	0.27	0.81	0.41
Influenza	246	0.72	0.83	0.77
Other	312	0.79	0.51	0.62
None	6	0.32	1.00	0.48
Average	744	0.68	0.61	0.62

Table 5: Pertinence classification results

The attribute classifier is a support vector machine (SVM) trained with stochastic gradient descent using scikit-learn (Pedregosa et al., 2011). Each entity is represented as the average word embedding, concatenated with the word embeddings for the previous and next 5 words. We also include the speaker code of the utterance in which the entity appears. We train the model on 252 conversations and test on 50 for which we have human-assigned modality and pertinence labels.

We classify entities into the following modality categories: actual, negative, possible, or none. Table 4 shows the results of modality classification on the test set of 50 conversations. Since the majority of entities have a modality of 'actual', the model performs the best on this class. Entities are also classified as pertinent to one of the disease categories, or none. Table 5 shows the results of pertinence classification. Again we see that the classifier performs the best on the classes with more examples.

Modality classification performs fairly well with a context window of 5, likely because the relevant information can be found nearby in the text. However, pertinence classification is not as accu-

`index.html`

[9] `https://www.nlm.nih.gov/research/umls/rxnorm/`

rate, perhaps because it requires more global information about what conditions the patient has. In some cases, pertinence may be purely determined by a clinicians medical knowledge, not the information present in the text.

In the future we hope to have more annotated data on which to train, which should improve the overall performance, especially for the smaller classes.

5.4 Clinical note generation

In the note generation phase, we convert the structured data from the previous steps (i.e., entities and their attributes) into a free text clinical note that resembles what a physician would have written. This involves organizing the entities according to a structured note organization and, finally, generating the text of the note.

5.4.1 SOAP entity classification

After extracting clinical entities, we classify them according to the traditional four sections of a clinical note: subjective (S), objective (O), assessment (A), plan (P) (Bickley and Szilagyi, 2013). We also add a 'none' category, which means that the given entity should not be included in the note.

The SOAP classifier is a neural network trained on each word of the entity, the previous and next five words, the speaker code of the corresponding utterance, and the type of entity. The text and context are represented as word embeddings using the PubMed word2vec model. Since the note generation requires special annotations, we currently only have 58 conversations for training, and 20 for test.

Table 6 shows the results of SOAP classification. The model is the most accurate at determining which information to exclude from the note.

Type	Instances	P	R	F_1
S	299	0.52	0.56	0.54
O	51	0.44	0.43	0.44
A	55	0.35	0.16	0.22
P	66	0.22	0.15	0.18
None	708	0.69	0.72	0.70
Average	1189	0.59	0.61	0.60

Table 6: SOAP classification results

Class	P	R	F_1
ADHD	0.84	0.84	0.83± 0.05
Depression	0.80	0.64	0.71 ± 0.08
Osteoporosis	0.81	0.78	0.78 ± 0.04
Influenza	0.91	0.95	0.93 ± 0.04
COPD	0.75	0.65	0.68 ± 0.14
Type II Diabetes	0.81	0.75	0.76± 0.05
Other	0.71	0.82	0.76± 0.05
Average	0.79	0.78	0.78± 0.04

Table 7: Primary diagnosis classification results. 800 dyads using 5-fold cross-validation (Train: 80%, Test: 20%). F_1 score is the mean ± variance.

5.4.2 SOAP note generation

Our current note generation step organizes the entities into the SOAP sections, and lists them along with their attributes. Actually generating full sentences that more closely resemble a physician-generated note is the next step for our future work.

5.5 Primary diagnosis classification

We classify the primary diagnosis for each conversation. The purpose of this task is to automatically identify the main diagnosis for billing codes. We train and test the models on a 5-fold cross validation of the 800 dyads. We apply *tf-idf* on the cleaned text of each dyad and then use logistic regression, SVMs with various parameter settings, and random forest models for classification. The F_1 score is calculated based on the human-assigned labels available in the transcription.

The primary diagnosis classifier performs reasonably well even without labeled entity features. The results for influenza achieve almost 0.90 F_1 score, while the results for COPD and depression obtain an F_1 score of approximately 0.70. By inspecting the conversations, we find that visits with a primary diagnosis of depression mostly consist of general discussions related to daily routine, family life, and mood changes, which often result in misclassification probably because no medical terms are mentioned. By contrast, in patient visits where the primary diagnosis is influenza, the discussion is more focused on the disease.

The top words used by the classifier were *H1N1, ache, temperature, sore, sick, symptom, swine, body,* and *strep*, which possibly makes it easier to classify. On the other hand, COPD is misclassified mostly as the category 'other', which includes diseases such as asthma, CHF (Congestive heart failure), hypercholesterolemia, atopic dermatitis, HIV/AIDS, prenatal visit, hypercholesterolemia. That is, the COPD dyads may be misclassified because of the presence of other respiratory diseases in the 'other' category. We plan to extend the diagnosis classifier to multi-label classification.

5.6 Topic modeling

Topic modeling is an unsupervised machine learning technique used to form k topics (i.e., clusters of words) occurring together, where k is usually chosen empirically. We perform topic modeling with LDA using the open-source gensim package (Řehůřek and Sojka, 2010) with varying numbers of topics k =(5, 10, 12, 15, 20, 25, 30, and 40).

Due to their colloquial nature, patient-clinician conversations contain many informal words and non-medical terms. We remove common words, including stop words from NLTK (Steven Bird and Loper, 2009), backchannel words (like '*uh-huh*'), and words with frequencies above 0.05% of the total number words in all the documents (to reduce the influence of more generic words).

The topic modeling results are shown in Table 8; we choose k=12 topics because they provided the best topic distribution and coherence score. The words in each topic are reported in decreasing order of importance.

A manual analysis shows that topic 0 captures words related to ADHD and depression, while topic 1 is related to asthma and flu, and topic 3 is related to women's health, and so on. This qualitative evaluation of topics shows that topic modeling can be helpful in extracting important information and identifying the dominant topic of a conversation. In our future work, we also plan to do a quantitative evaluation of topic modeling results using state-of-the-art methods such as the methodology proposed by Wallach et al. (2009).

We see the potential use of topic modeling to keep track of the focus of each visit, the distribution of word usage, categorization, and to group patients together using similarity measures. We also use it for relevant text extraction in the next section.

5.7 Relevant utterance extraction

Identifying the key parts of the doctor-patient conversation can be helpful in finding the relevant information. In the previous section, we observe that topic modeling can be helpful in identifying the underlying topics of the dyads. We also use topic modeling to extract the utterances relevant to the primary disease diagnosis.

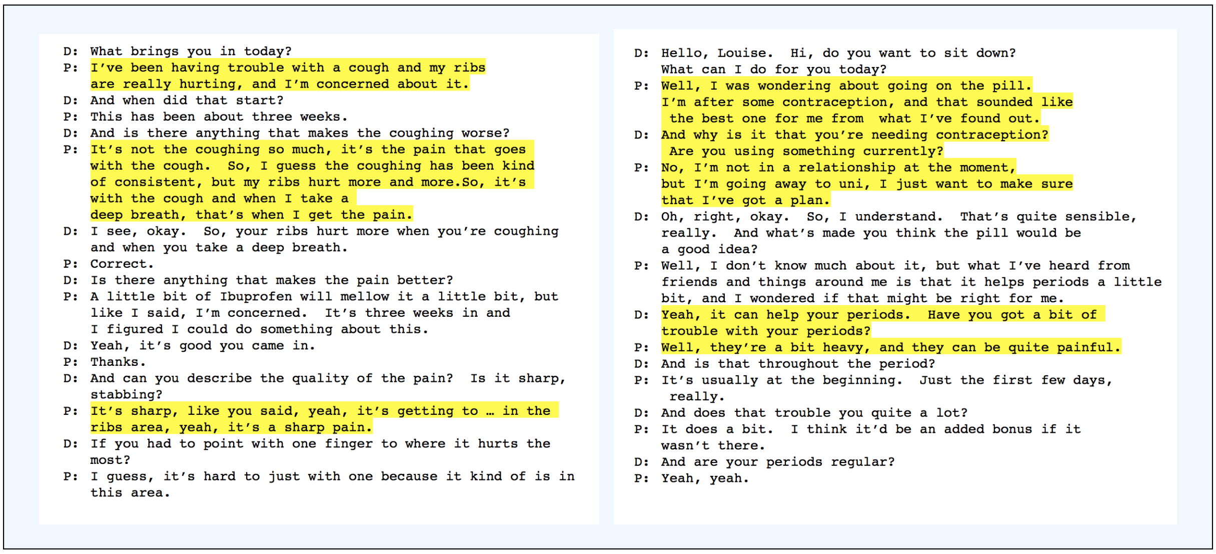
Topic#	Topic words
0	focus, sleeping, depressed, asleep, attention, mind, cymbalta, appetite, psychiatrist, energy
1	ache, h1n1, treat, asthma, temperature, diarrhea, anybody, mucinex, chill, allergic
2	period, knee, birth, heavy, ultrasound, iron, metoprolol, pregnancy, pregnant, history,
3	meal, diabetic, lose, unit, mail, deal, crazy, card, swelling, pound
4	cymbalta, lantus, cool, cancer, crazy, allergy, sister, attack, nurse, wow
5	referral, trazodone, asked, shingle, woman, medicare, med, friend, clinic, form
6	breo, cream, puff, rash, smoking, albuterol, skin, allergy, proair, allergic
7	fosamax, allergy, tramadol, covered, plan, calcium, bladder, kept, alcohol, ache
8	metformin, x-ray, nerve, knee, lasix, bottle, lantus, hurting, referral, switch
9	lantus, looked, injection, botox, changed, flare, happening, cream, salt, sweating
10	generic, triumeq, cost, farxiga, physical, therapy, gosh, fracture, increase, invokana
11	unit, list, appreciate, therapy, difference, counter, report, lasix, lantus, endocrinologist

Table 8: Topic Modeling: Top 10 words for 12 topics.

We apply the following steps adapted from a publicly available text summarization method[10]:

1. Fit the LDA model to all dyads.
2. Pass the dyads for each class to the LDA model to determine the class-wise topic distribution.
3. Select the dominant topics for each class using the topic weight matrix.
4. For each dyad within this subset:

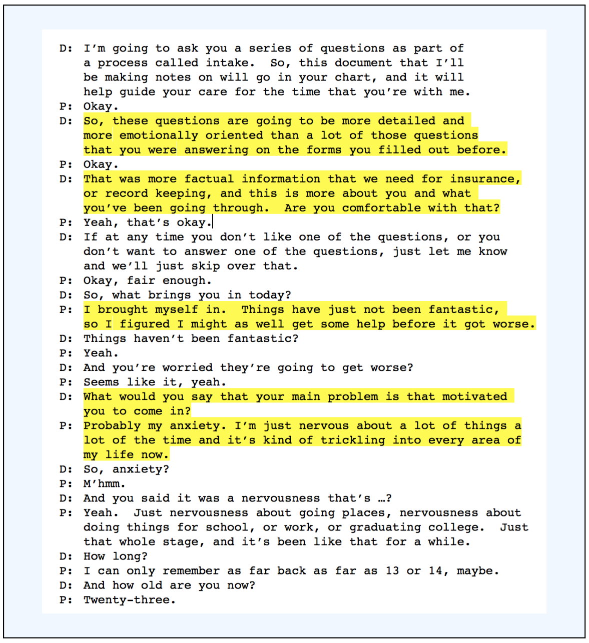
[10] https://github.com/g-deoliveira/TextSummarization



```
D: What brings you in today?
P: I've been having trouble with a cough and my ribs
   are really hurting, and I'm concerned about it.
D: And when did that start?
P: This has been about three weeks.
D: And is there anything that makes the coughing worse?
P: It's not the coughing so much, it's the pain that goes
   with the cough.  So, I guess the coughing has been kind
   of consistent, but my ribs hurt more and more.So, it's
   with the cough and when I take a
   deep breath, that's when I get the pain.
D: I see, okay.  So, your ribs hurt more when you're coughing
   and when you take a deep breath.
P: Correct.
D: Is there anything that makes the pain better?
P: A little bit of Ibuprofen will mellow it a little bit, but
   like I said, I'm concerned.  It's three weeks in and
   I figured I could do something about this.
D: Yeah, it's good you came in.
P: Thanks.
D: And can you describe the quality of the pain?  Is it sharp,
   stabbing?
P: It's sharp, like you said, yeah, it's getting to … in the
   ribs area, yeah, it's a sharp pain.
D: If you had to point with one finger to where it hurts the
   most?
P: I guess, it's hard to just with one because it kind of is in
   this area.
```

```
D: Hello, Louise.  Hi, do you want to sit down?
   What can I do for you today?
P: Well, I was wondering about going on the pill.
   I'm after some contraception, and that sounded like
    the best one for me from  what I've found out.
D: And why is it that you're needing contraception?
    Are you using something currently?
P: No, I'm not in a relationship at the moment,
   but I'm going away to uni, I just want to make sure
   that I've got a plan.
D: Oh, right, okay.  So, I understand.  That's quite sensible,
   really.  And what's made you think the pill would be
   a good idea?
P: Well, I don't know much about it, but what I've heard from
   friends and things around me is that it helps periods a little
   bit, and I wondered if that might be right for me.
D: Yeah, it can help your periods.  Have you got a bit of
   trouble with your periods?
P: Well, they're a bit heavy, and they can be quite painful.
D: And is that throughout the period?
P: It's usually at the beginning.  Just the first few days,
    really.
D: And does that trouble you quite a lot?
P: It does a bit.  I think it'd be an added bonus if it
   wasn't there.
D: And are your periods regular?
P: Yeah, yeah.
```

Figure 6: (Left) Presenting problem: *Cough and rib pain.* (Right) Presenting problem: *Women's health and contraception.* Extracted utterances are highlighted.



```
D: I'm going to ask you a series of questions as part of
   a process called intake.  So, this document that I'll
   be making notes on will go in your chart, and it will
   help guide your care for the time that you're with me.
P: Okay.
D: So, these questions are going to be more detailed and
   more emotionally oriented than a lot of those questions
   that you were answering on the forms you filled out before.
P: Okay.
D: That was more factual information that we need for insurance,
   or record keeping, and this is more about you and what
   you've been going through.  Are you comfortable with that?
P: Yeah, that's okay.
D: If at any time you don't like one of the questions, or you
   don't want to answer one of the questions, just let me know
   and we'll just skip over that.
P: Okay, fair enough.
D: So, what brings you in today?
P: I brought myself in.  Things have just not been fantastic,
   so I figured I might as well get some help before it got worse.
D: Things haven't been fantastic?
P: Yeah.
D: And you're worried they're going to get worse?
P: Seems like it, yeah.
D: What would you say that your main problem is that motivated
   you to come in?
P: Probably my anxiety. I'm just nervous about a lot of things a
   lot of the time and it's kind of trickling into every area of
   my life now.
D: So, anxiety?
P: M'hmm.
D: And you said it was a nervousness that's …?
P: Yeah.  Just nervousness about going places, nervousness about
   doing things for school, or work, or graduating college.  Just
   that whole stage, and it's been like that for a while.
D: How long?
P: I can only remember as far back as far as 13 or 14, maybe.
D: And how old are you now?
P: Twenty-three.
```

Figure 7: Presenting problem: *Anxiety*. Extracted utterances are highlighted.

(a) Split the conversation into sentences, using the NLTK (Steven Bird and Loper, 2009) sentence tokenizer.

(b) Determine the topic distribution of each sentence using LDA.

(c) Filter out the sentences whose dominant topic is not equal to the dominant topic of that dyad. What is left is a subset of sentences that reflect the given topic.

We conduct experiments on all 800 dyads and the 11 dyads from YouTube. Topic modeling is performed exactly as described in the previous section, with 12 topics. The results are shown in Table 6 and 7. The three dyads shown are from open-source YouTube data focusing on (a) cough and rib pain, (b) women's health and contraception, and (c) anxiety, respectively.

The results indicate a reasonable quality of relevant text extraction despite the limited amount of data. We can see that many of the utterances discussing the presenting problem are extracted. Since we do not have labels for the true relevance of the sentences to the disease, we are unable to provide any quantitative metrics, which is the subject of future work.

6 Conclusion & future work

The cumulative output of these models constitutes the initial automated system. Although for these experiments we used manual transcriptions, in practice the input would be from automatic speech recognition (ASR). Future research will include using ASR to record transcripts in real time, as well as expanding the types of entities we extract, identifying quantity, quality, and severity.

Diagnosis classification currently handles 6 classes only, and does not account for conditions other than the primary diagnosis that may be discussed in the conversation. We will also expand diagnosis classification to handle more classes, and to predict multiple diagnoses.

We have presented a system for extracting clinically relevant entities from physician-patient dialogues using linguistic context. The results show that clinical note-taking can be at least partially automated, saving clinicians valuable time. This system can result in a streamlined data entry process and a cleaner EMR note that can be used for analytics and automated decision making.

References

Moumita Bhattacharya, Claudine Jurkovitz, and Hagit Shatkay. 2017. Identifying patterns of co-occurring medical conditions through topic models of electronic health records. In *AMIA, iHealth 2017 Clinical Informatics Conference*.

Lynn S. Bickley and Peter G. Szilagyi. 2013. *Bates' pocket guide to physical examination and history taking 7th ed.* Wolters Kluwer Health/Lippincott Williams Wilkins.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format Switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation 2010*, 44:387–419.

Katherine Redfield Chan, Xinghua Lou, Theofanis Karaletsos, Christopher Crosbie, Stuart Gardos, David Artz, and Gunnar Rätsch. 2013. An empirical analysis of topic modeling for mining cancer clinical notes. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 56–63. IEEE.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Longxu Dou, Guanghui Qin, Jinpeng Wang, Jin-Ge Yao, and Chin-Yew Lin. 2018. Data2Text studio: Automated text generation from structured data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 13–18, Brussels, Belgium. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology, Chapter 11.* Sage, Beverly Hills, CA, USA.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Peter J Liu. 2018. Learning to write notes in electronic health records. *ArXiv eprint 1808.02622v1*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of NAACL-HLT 2018*, pages 1101–1111.

Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2018. Deepr: A convolutional net for medical records. *IEEE Journal of Biomedical and Health Informatics*, 21:22–30.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*, pages 2227–2237.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. 2018. Scalable and accurate deep learning for electronic health records. *NPJ Digital Medicine*, 2018:1–10.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Carol P Roth, Yee-Wei Lim, Joshua M. Pevnick, Steven M. Asch, and Elizabeth A. McGlynn. 2003. The challenge of measuring quality of care from the electronic health record. *American Journal of Medical Quality*, 24:385–394.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165:753–760.

Ewan Klein Steven Bird and Edward Loper. 2009. *Natural Language Processing with Python*. OReilly Media.

Jannik Strötgen and Michael Gertz. 2010. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation 2013*, 47:269–298.

Krish Thiru, Alan Hassey, and Frank Sullivan. 2003. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ*, 326:1070–1072.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112. ACM.

Nicole Gray Weiskopf and Chunhua Weng. 2013. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20:144–151.