AutoScribe: Extracting clinically pertinent information from patient-clinician dialogues

Faiza Khan Khattak*ab, Serena Jebleeebh, Noah Cramptonc, Muhammad Mamdani†, Frank Rudicza,b,c,d

*Department of Computer Science, University of Toronto, Toronto, Ontario, Canada (†equal contribution)
†Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada
‡Li Ka Shing Knowledge Institute, St Michael’s Hospital, Toronto, Ontario, Canada
§Surgical Safety Technologies Inc, Toronto, Ontario, Canada

Abstract
We present AutoScribe, a system for automatically extracting pertinent medical information from dialogues between clinicians and patients. AutoScribe parses the dialogue and extracts entities such as medications and symptoms, using context to predict which entities are relevant, and automatically generates a patient note and primary diagnosis.

Keywords:
Medical Records, Machine Learning, Medical Informatics

Introduction
Currently, clinicians spend up to 50% of their time entering information from patient interviews into electronic medical records (EMRs) [1]. This reliance on slow, laborious, and inconsistent data entry results in wide variability in the quality of EMR data [2], which presents a challenge to clinical data analytics [3]. Recent machine learning (ML) algorithms, such as recurrent neural networks and word embeddings [4], have been applied to tasks such as disease and mortality prediction from EMR data [5,6]. This suggests that a significant portion of clinical data entry can be automated by analyzing patient-clinician dialogues.

Here, we optimize an ML model, AutoScribe, to classify dialogue phrases from patient interviews as contextually pertinent to clinical documentation, which is the foundational step to generating EMR data from the analysis of patient-clinician dialogues. We extract medically relevant entities such as signs, symptoms, diagnoses, therapies, and referrals through natural language processing. Unlike systems which primarily use lexicon-based term matching, our system also uses linguistic context and time information.

Data
The data consists of 800 audio patient-clinician dialogues and their transcripts, purchased from Verilogue Inc1, including primary diagnosis codes. The most frequent are ADHD, COPD, depression, and influenza.

We developed a new annotation tool and are doubly annotating all dyads for relevant medical entities. Of the 30 dialogues that have been completed, the annotations have .53 agreement (Krippendorff’s alpha [8]) and .80 partial match F1 score. We also have 302 dialogues with annotations from one physician at present. We present a synthetic patient-clinician dialogue in Table 1 with the output of our system compared to human annotation.

Methods and Results

1 http://www.verilogue.com
2 http://bio.nlplab.org/
3 http://groups.inf.ed.ac.uk/ami/corpus/
4 https://bioportal.bioontology.org/ontologies
5 http://consumerhealthvocab.chpe.utah.edu/CHVwiki/
6 http://www.snomed.org/
7 https://www.nlm.nih.gov/research/umls/rxnorm/
Table 1 – Example dialogue - (a) Human annotation. (b) Automatic annotation. In both 1a and 1b, highlight indicates the annotated entities; darker highlights indicate overlap between human and automatic annotations. Subscripts indicate the entity type (TIMEX3 indicates time phrases).

| DR: Okay, that’s good. Let’s keep you on the [same dose of | Metformin] Medication [for now] TIMEX3 then we’ll check your [a1c] Investigation/Therapy again [in three months] TIMEX3 , and then I’ll [see you back here after that] Disposition plan. |
| DR: Okay, that’s good. Let’s keep you on the same [dose] Medication of [Metformin] [medication for now] TIMEX3 then we’ll check your [a1c] again [in three months] TIMEX3 , and then I’ll see you back here after that. |

Attribute classification

Once the entities have been identified, the system should determine which are actually pertinent to the diagnosis. For instance, a physician or patient might mention a medication that they have never actually taken, so the system should not record that medication as part of the patient’s history. Currently, we classify two attributes: modality and pertinence. The modality indicates whether the event actually occurred (actual, negative, possible), and pertinence indicates the condition to which the entity is medically relevant (i.e., ADHD, COPD, depression, influenza, other). The attribute classifier is a support vector machine (SVM) trained with stochastic gradient descent [9].

Each medical 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. The system achieves .77 F1 score for modality classification, and .62 for pertinence. Pertinence classification currently performs worse than modality, perhaps because it requires more global information.

Primary diagnosis classification

We classify the primary diagnosis on each patient-clinician conversation to be used 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 use logistic regression, SVMs, and random forest models. The F1 scores of classification are calculated based on the human-assigned labels available in the transcription of the conversation’s ‘primary diagnosis’ field. 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.

F1 scores (Linear SVM): Influenza .93 ± .04, ADHD .83 ± .05, COPD .68 ± .14, Osteoporosis .78 ± .04, Type II diabetes .76 ± .07, Depression .71 ± .08, and Other .76 ± .05.

Discussion & Conclusion

We have presented a novel approach to clinician-patient dialogue parsing, whose outputs are oriented toward pragmatic linguistic features, and the needs of clinicians. Specifically, we have developed machine learning models based on recurrent neural networks that extract medical linguistic entities and their time-based contextual partners, as well as primary diagnoses from dialogue. Future directions include extracting other key contextual entities within clinical dialogues that are pertinent to clinical documentation, such as quantity, quality, and severity words and phrases, as well as accounting for similar medical terms and spelling variations. Training will be expanded to include more entities, more conversations, more diagnoses, and multiple diagnoses per conversation.

Acknowledgments

Research Ethics Board approval from St Michael’s Hospital (REB # 18-082) and the University of Toronto (REB# 00036367).

References


Address for correspondence