AutoScribe: Extracting clinically pertinent information from patient-clinician dialogues

Faiza Khan Khattak[a,b*], Serena Jeblee[a,b*], Noah Crampton[c], Muhammad Mamdani[c], Frank Rudzicz[a,b,c,d]

*[a] Department of Computer Science, University of Toronto, Toronto, Ontario, Canada (*equal contribution)*
[b] Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada
[c] Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Ontario, Canada
[d] 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 Inc[1], 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 F_1 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

The AutoScribe system currently consists of several modules. The cumulative output of these models constitutes the initial AutoScribe system. We evaluate each component of the system using F_1 measure, considering tags that overlap with the human annotation as correct. For entity tagging, we also calculate inter-annotator agreement between the physicians and the automatic pipeline using Krippendorff's alpha [8]. All but the utterance type classification model are evaluated on 302 conversations.

Utterance type classification

Each utterance in the dialogue is automatically labeled as a *question, statement, positive answer, negative answer, backchannel* or *excluded.* We use a two-layer bidirectional gated recurrent unit (GRU) neural network [12], implemented in PyTorch. Each word is represented as a 200-dimensional vector using the freely available Wikipedia-PubMed word embedding model[2]. We evaluate the utterance type classifier on 20 conversations, annotated independently by 2 annotators with inter-annotator agreement of .77 (Cohen's kappa). For training, we use two external, publicly available datasets: the Switchboard corpus [10], and the AMI corpus[3]. Our model achieves .71 F_1 score on Verilogue data.

Time expression identification

Phrases in the dialogue that reference absolute and relative times and dates are automatically tagged and converted to standardized values using HeidelTime [11], a freely available temporal tagger. For example, in a document dated Jan 1, 2018, the phrase *tomorrow* would be normalized to *2018-01-02*.

Medical entity identification

AutoScribe currently identifies the following medical concepts: anatomical locations, signs and symptoms, diagnoses, medications, referrals, investigations and therapies, and reasons for visit. The identification uses lexicon look-up using terms from BioPortal[4], Consumer Health Vocabulary (CHV)[5], SNOMED-CT[6], and RxNorm[7], and achieves an average F_1 score of .63 and .55 Krippendorff's alpha. 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.

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.chpc.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: *How's the [numbness in your toes]*$_{\text{Sign/Symptom}}$ */[toes]*$_{\text{Anatomical Location}}$ *?* ***PT:*** *The same. I'm used to it by now.* ***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]*$_{\text{Disposition plan}}$*.*	***DR:*** *How's the numbness in your [toes]*$_{\text{Anatomical Location}}$ *?* ***PT:*** *The same. I'm used to it by [now]*$_{\text{TIMEX3}}$*.* ***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.*

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 F_1 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 F_1 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.

F_1 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

[1] Sinsky C. et al. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Ann Intern Med.* 2016 Dec 6; 165 (11): 753-760. doi: 10.7326/M16-0961

[2] Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *Br Med J*. 2003; 326(7398): 1070–2.

[3] Roth CP, et al. The challenge of measuring quality of care from the electronic health record. *American Journal of Medical Quality* 2009; 24(5): 385–94. http://dx.doi.org/10.1177/1062860609336627. PMid:19482968.

[4] Mikolov T, Sutskever I, Chen K, Corrado G, and Dean J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (NIPS'13) 2013; 3111-3119.

[5] Rajkomar A, Oren E, Chen, K, Dai AM, Hajaj N, Liu PJ, … Dean J. (2018). Scalable and accurate deep learning for electronic health records. NPJ Digital Medicine (2018): 1–10. https://doi.org/10.1038/s41746-018-0029-1

[6] Mullenbach J, Wiegreffe S, Duke J, Sun J, and Eisenstein J. Explainable prediction of medical codes from clinical text. In *Proceedings of NAACL-HLT* 2018; 1101–1111.

[7] Liu PJ. Learning to write notes in electronic health records. *ArXiv preprint 1808.02622v1*, 2018.

[8] Krippendorff K. Content Analysis: An Introduction to Its Methodology, Chapter 11. Sage, Beverly Hills, CA, 2004.

[9] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, and Bengio Y. 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)* (2014): 1724–1734.

[10] Calhoun S, Carletta J, Brenier J, Mayo N, Jurafsky D, Steedman M, and Beaver D. The NXT-format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue. *Language Resources and Evaluation* 2010; 44(4), 387-419, doi:10.1007/s10579-010-9120-1.

[11] Strötgen J and Gertz M. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation* 2013; 47(2):269–298.

[12] Pedregosa F, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011; 12: 2825-2830.

Address for correspondence

Faiza Khan Khattak. Email: faizakk@cs.toronto.edu