ISSN No:-2456-2165

Infusing Machine Learning and Computational Linguistics into Clinical Notes

Funke V. Alabi.
Nova South eastern University,
Florida, USA.

Onyeka Omose
University of Denver, Colorado,
USA.

Omotomilola Jegede.
Old Dominion University Norfolk,
Virginia, USA.

Abstract:- Entering free-form text notes into Electronic Health Records (EHR) systems takes a lot of time from clinicians. A large portion of this paper work is viewed as a burden, which cuts into the amount of time doctors spend with patients and increases the risk of burnout. We will see how machine learning and computational linguistics can be infused in the processing of taking clinical notes. We are presenting a new language modeling task that predicts the content of notes conditioned on historical data from a patient's medical record, such as patient demographics, lab results, medications, and previous notes, with the goal of enabling AI-assisted note-writing. Using the publicly available, de-identified MIMIC-III dataset, we will train generative models and perform multiple measures of comparison between the generated notes and the dataset.We will have detailed discussionabouthow thesemodels can help with assistivenote-writing functions like auto- complete and error-detection.

I. INTRODUCTION

Astudy (Sinsky et al., 2016) found that for every hour spent in person with patients, doctors spend almost two hours on administrative tasks. Inputting clinical notes into the electronic health record (EHR) software, which records patient information such as health history, assessment (such as diagnoses), and treatment plan, takes the longest among administrative tasks. Asignificant portion of the documentation requirements are seen as tedious, which are a major cause of burnout and career dissatisfaction for clinicians. Moreover, when this task takes up more of the patients' time, it affects how satisfied they are with their experience.

Due to the importance of this problem, various industries have emerged to offer support for documentation. These include services like dictation, where medical professionals verbally provide notes for transcription by either a human or computer backend, and scribes, who are human assistants specifically tasked with recording notes. In agreement with Gellert et al. (2015), we interpret this as evidence that the usability of EHR software needs to be enhanced. Languagemodels are useful in assistive writing features for notes, like auto-completion and error-checking. Such features would probably be more effective if model used is stronger. Therefore, developing language models for clinical notes is the main goal of this work.

The de-identified electronic health records of 39,597 patients from an extensive tertiary care hospital's intensive-care unit (ICU) makeup the MIMIC-III dataset. This is the most comprehensive publicly available dataset of its kind, containing information on patient demographics, ordered medications, laboratory measurements, and—most importantly for this work—notes recorded by healthcare providers. Numerous studies, primarily aimed at predicting clinical events such as acute kidney injury (Mandelbaum et al., 2011), mortality (Johnson et al., 2017), or diagnoses and medication orders (Choi et al., 2016), have been sparked by the release of the data and its predecessor MIMIC-II (Saeed et al., 2011).Alot of EHR datasets aremostly private or lack clinical notes, however they have been used in the past to forecast clinical events.

A great deal of previous research has been done on the use of clinical notes for various purposes. Using structured output extracted from notes, Friedman et al. (2004) created Unified Medical Language System (UMLS) codes. Frequently, notes are utilized as input for machine learning models that forecast clinically relevant events in the future based on historical electronic health record data (Miotto et al. (2017), Rajkomar et al. (2018)). Using expert rules and templates, Portetet al. (2009) automatically summarized EHR ICU data into text format. While they do not use data from the EHR, Jing et al. (2017) train models to generate medical imaging reports from x-rays, a type of image captioning task.

Using a class of methods based on sequence-to-sequence learning, conditional language modeling has also been thoroughly studied, including for machine translation (Wu et al., 2016) and speech recognition (Chiu et al., 2017).There, we model an input sequence (French words, for example) that is conditioned on an output sequence (English words). However, the majority of earlierresearch relies on text-to-text or audio-to-text mapping, which maps one modality to text. The data used in our work is conditioned on a wide range of sequential structured and unstructured (text) data, as well as static and time-dependent data.

II. MATERIALS AND METHODS

A. *Conditional Language Models for Clinical Notes*

Language models represent a sequence of tokens (e.g., words), w1, w2,...,wn, that specify a probability distribution, P(w1, w2,..., wn), over language pieces (e.g., sentences or documents).

We can write using the chain-rule of probability.

$$P(w_1, w_2, ... w_n) = \prod_{i=1}^{n} P(w_i | w_1, ..., w_{i-1})$$

which breaks down the joint distribution into a product of current token conditional distributions given past tokens. Additionally, this defines age nerative model that enables us to generate complete sequences by sampling likely tokens one at a time.

Conditional language models are comparable, with the exception that they receive extra context information, c:

$$P(w_1, w_2, \dots, w_n | c) = \prod_{i=1}^{n} P(w_i | c, w_1, \dots, w_{i-1})$$

Any media, including text, audio, images, and so on, could be the context. For instance, in machine translation, the sentence to be translated in the source language is this context.An image may serve as the context in picture captioning. In this work, the sequence to be predicted is the text of the current clinical note and context is the past data extracted from the Electronic Health Record (EHR) for the same patient, R. We also augment the context with the intended note type, T, and a hint of 10 tokens from the current note, H. This note-context, c = (R, T, H), is intended to simulate what is known at the time a clinician begins writing a note.

Formally, we develop the model:

$$P(D = w_1, w_2, \dots, w_n | c = (R, T, H))$$

Where HL=10,which is less than n, and H=w1,w2,...,wHL. We set n=500asthemaximumnote tokens to predict. We can see this as a supervised problem in machine learning, where we need to learn the mapping c → D.

B. Extracting and Representing Context

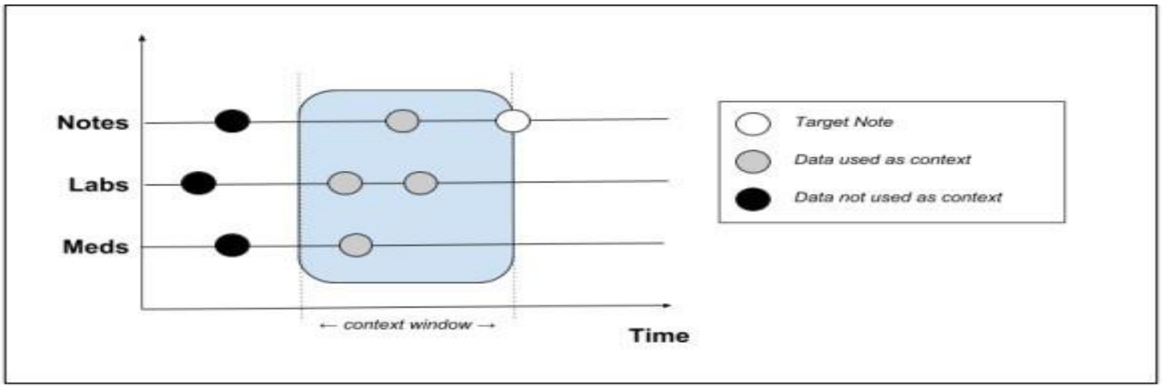


Fig. 1: Schematic showing which context data is extracted from the patient record

As seen in Figure 1, in our experiments, context data is only taken into account for the 24 hours preceding the note's writing time.

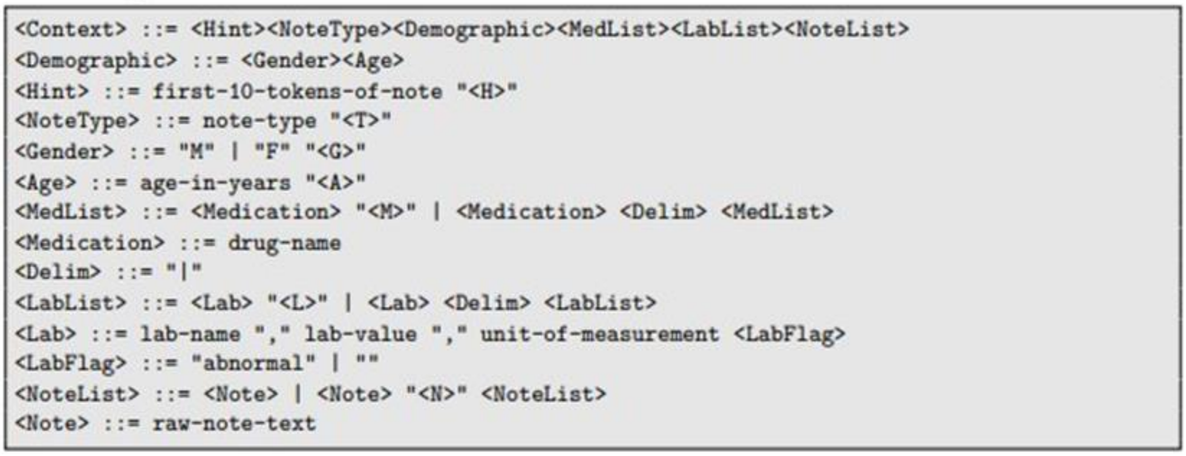
Asizable, publicly accessible database called MIMIC-III contains de-identified medical records linked to more than forty thousand patients who visited Beth Israel Deaconess Medical Center critical care units between 2001 and 2012. Data from laboratory tests, procedures, medications, vital sign measurements taken at the bedside (less than one data point per hour), and demographics are all included in the database(Johnson et al., 2017).

We experiment with the following context data classes:

- **Demographic data(D):** This is static data for each patient in MIMIC-III, and it can be found in the Patients table. Using the date-of-birth field, we determine the age at the time of the note and extract the gender.
- **Medications (M):** We take the names of the medications that the patient has been prescribed from the Prescriptions table. These are prescription drugs that were given during the context window.
- **Labs (L):** We retrieve the lab name, value, unit of measurement, and, if available, the flag indicating whether the data is abnormal from the Lab Events table. These lab tests were requested within the context window.
- **Past notes (Np):** The context window is used to display the text of past notes.

The Appendix A.1 andA.2 contains the a sample table from the data set and precise column names from the MIMIC-III table.

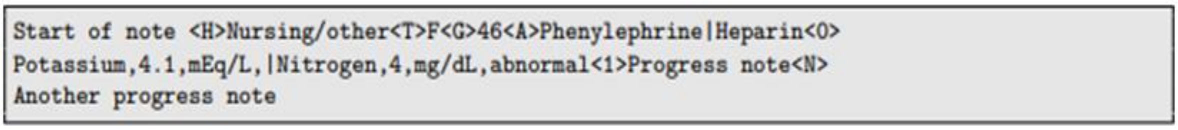
Theafore mentioned data elements are all transformed into a string representation using unique tokens that serve as data class separators. The Backus-Naur form notation is used to represent the context data as follows:



```
<Context> ::= <Hint><NoteType><Demographic><MedList><LabList><NoteList>
<Demographic> ::= <Gender><Age>
<Hint> ::= first-10-tokens-of-note "<H>"
<NoteType> ::= note-type "<T>"
<Gender> ::= "M" | "F" "<G>"
<Age> ::= age-in-years "<A>"
<MedList> ::= <Medication> "<M>" | <Medication> <Delim> <MedList>
<Medication> ::= drug-name
<Delim> ::= "|"
<LabList> ::= <Lab> "<L>" | <Lab> <Delim> <LabList>
<Lab> ::= lab-name "," lab-value "," unit-of-measurement <LabFlag>
<LabFlag> ::= "abnormal" | ""
<NoteList> ::= <Note> | <Note> "<N>" <NoteList>
<Note> ::= raw-note-text
```

Image A.1 – The Backus-Naur Form Notation

The following is an illustration of how the input is instantiated:



```
Start of note <H>Nursing/other<T>F<G>46<A>Phenylephrine|Heparin<0>
Potassium,4.1,mEq/L,|Nitrogen,4,mg/dL,abnormal<1>Progress note<N>
Another progress note
```

Image A.2 – The Backus-Naur Form Notation

We limit the quantity of context to kens in our experiments to 500.

C. Task Dataset Construction

Our supervised dataset is built in two phases. Each patient is initially assigned at random to either the training, validation, or test sets. In the second stage, we generate a supervised learning example, (c, N), for each note MIMIC-III, N. Since we want to use the model for any note pertaining to any patient, we did not perform any patient cohort filtering. To ensure that each patient appeared in exactlyoneset,wedividedthetrain,development,andtestsets.Thisisdonetostopthemodelfrom learning a patient's medical history during training and using that information to its advantage to predict exam questions.

The sizes of the train,validation, and test were, in order,1682869,201181,and198808.

D. Input and Target Encoding

Our task for conditional language modeling has been trimmed down to train a supervised model that maps note-text, N, to note-context, c = (R, H,T), a string sequence.The conversion of the data into input and output sequences is shown in Figure 2.

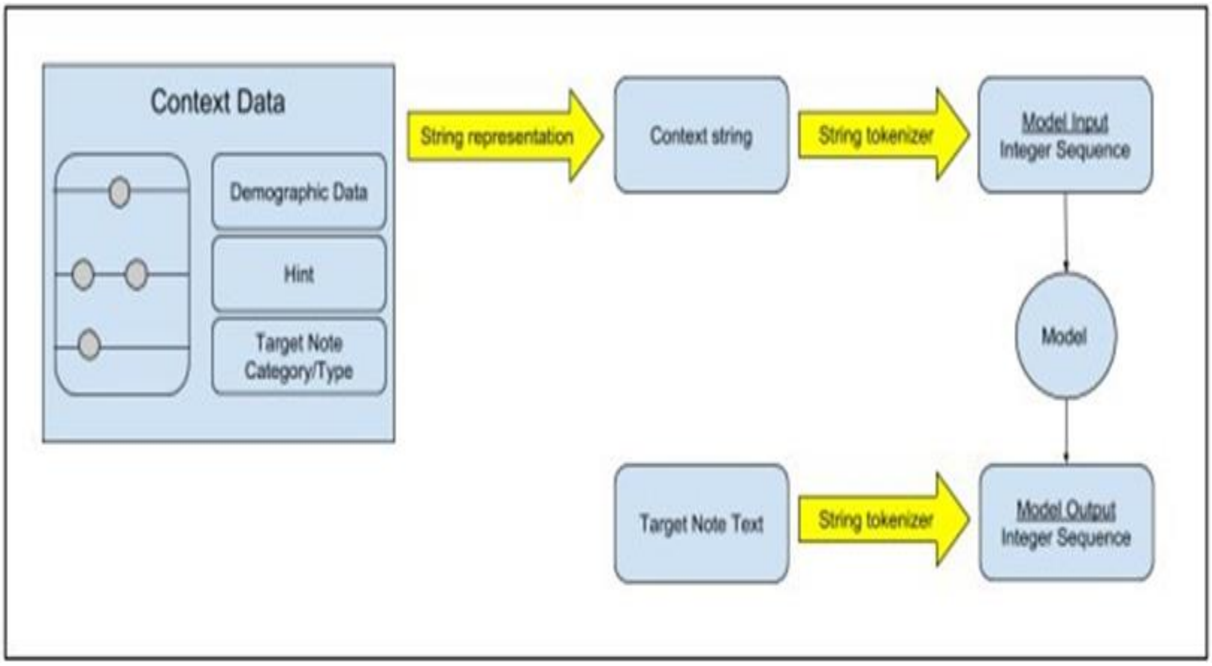


Fig. 2: Schematic showing how raw data is transformed to model training data

With a vocabulary of roughly 32,000 words, asub-word tokenizer is used to tokenize both theinput and the target. As a result, we are able to convert the input and output into integer sequences that represent sub word tokens from the same vocabulary. Specifically, we don't pre-process the note text and we keep all of the white space that was present in the initial raw notes. It has been observed that the presence of white space in clinical notes is crucial for segmenting information, and its absence would make the notes less readable.

E. Model Architectures

Our starting point for sequence-to-sequence learning is the Transformer architecture, which was first presented by Vaswani et al. (2017) and provides state-of-the-art machine translation performance. Unlike recurrent neural network techniques, which had dominated natural language processing up until its publication, this architecture is an encoder-decoder without any recurrent submodules. Compared to typical machine translation tasks where the input and output sequences are sentences, the input and output sequences in our task may be substantially longer. As a result, we also test a recently developed Transformer-based model known as Transformer with memory-compressed attention, or T-DMCA, which Liuetal. (2018) demonstrated to be useful for large scale (i.e., much longer sequences) abstractive summarization.

F. Evaluation

- *Evaluating the language models*

We assess our conditional language models using a variety of metrics, some common and some unique to our work:

- *Perplexity Pertoken(PPL)*

The perplexity per token, which measures how well the current token is predicted given previous tokens averaged over the entire sequence, is a standard intrinsic metric used in evaluating language models. Because it ignores the predictive power that exists beyond the next token, this measure is extremely local. Keep in mind that perlexities from models with various vocabularies cannot be truly compared. In this instance, every model has the same 32,000-word vocabulary.

Alog-perplexity is reported.

- *Accuracy of Next Token (ACC)*

The precision of forecasting the subsequent token based on the previous tokens.As with perplexity, this metric is very local.

- *ROUGE-1,ROUGE-2(R1,R2)*

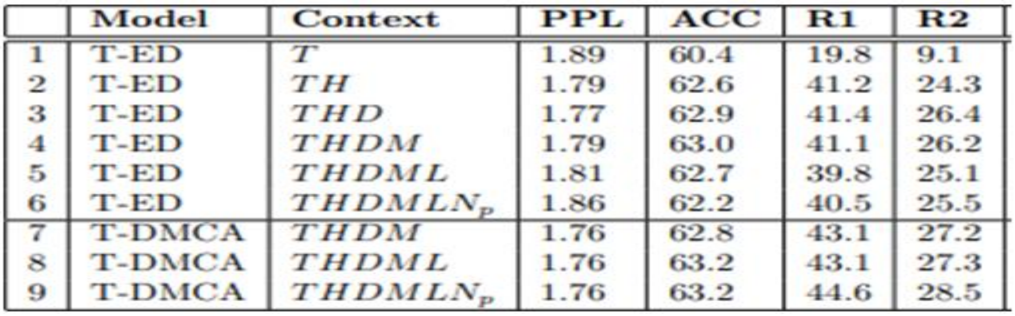
We examine n-gram recall and precision statistics comparing the candidate generated note and the ground truth note for a model-independent evaluation and a more comprehensive metric. To calculate ROUGE-1 (unigram) and ROUGE-2 (bigram) scores, we utilize the widely-used ROUGE package. The harmonic mean of ROUGE-P(precision) and ROUGE-R (recall), or the F1 variant, is what we report.

III. RESULTS AND DISCUSSION

A. Template Learning

We found that frequently used templates in the dataset were learned and consistently used in note generation during our analysis of model samples. This was observed even in the models trainedwith minimal data. A generated note (from the validation set) and the corresponding ground truth are displayed in Figure 3 of a model that is solely conditioned on note-type (T) and short hint (H). Despite the fact that a lot of the content has changed (for example, the patient table's age is in correct), the template was filled in with values that make sense for a radiology report and had suitable section titles and formatting. It was discovered that templates and styles particular to note types were also learned. The model does a good job of imitating the style and organization of the nursing note in Figure 4.

B. Quantitative Results



	Model	Context	PPL	ACC	R1	R2
1	T-ED	T	1.89	60.4	19.8	9.1
2	T-ED	TH	1.79	62.6	41.2	24.3
3	T-ED	THD	1.77	62.9	41.4	26.4
4	T-ED	$THDM$	1.79	63.0	41.1	26.2
5	T-ED	$THDML$	1.81	62.7	39.8	25.1
6	T-ED	$THDMLN_p$	1.86	62.2	40.5	25.5
7	T-DMCA	$THDM$	1.76	62.8	43.1	27.2
8	T-DMCA	$THDML$	1.76	63.2	43.1	27.3
9	T-DMCA	$THDMLN_p$	1.76	63.2	44.6	28.5

Table 1: Quantitative results for model architectures and HER contexts used in experiments

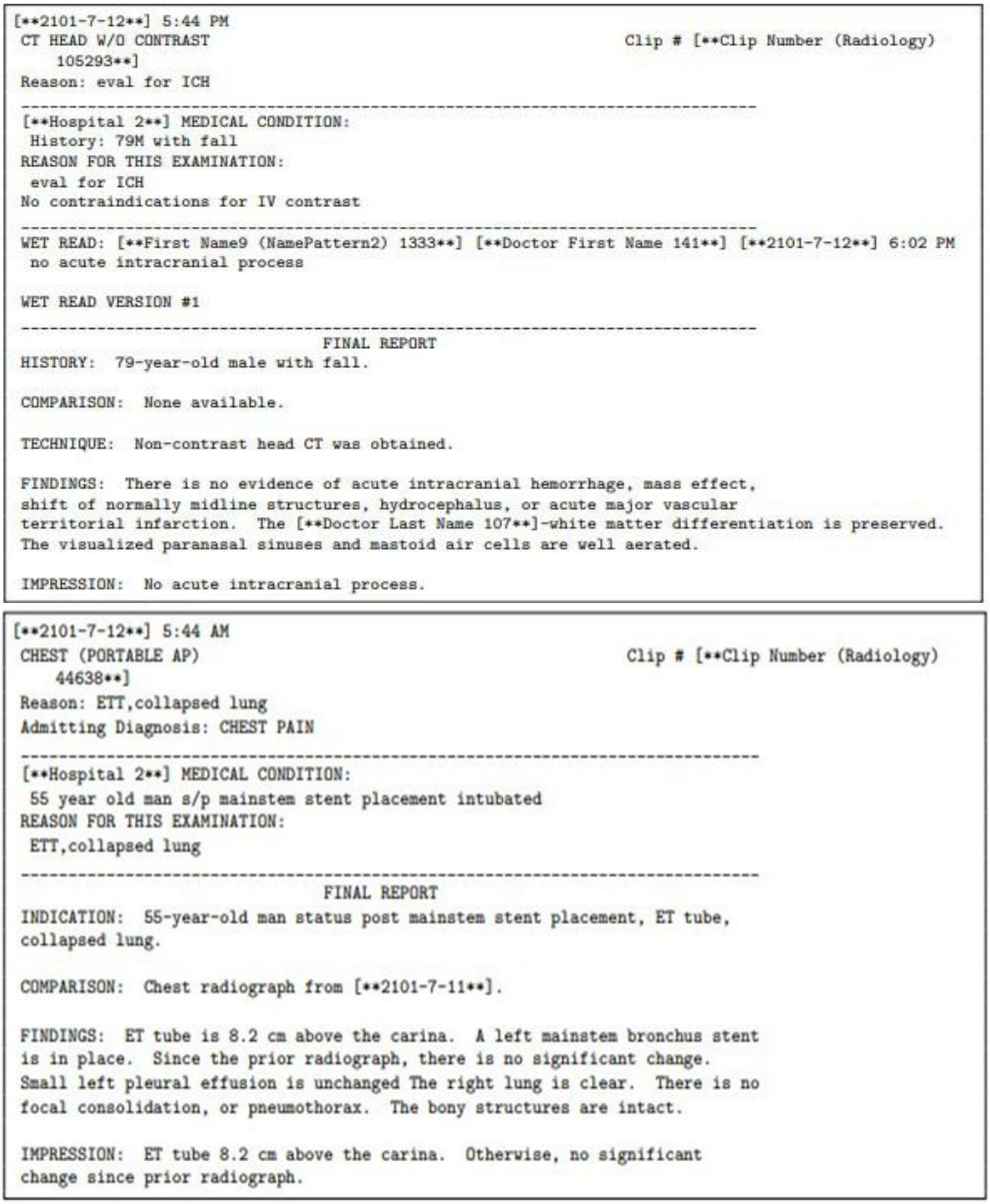
All the models trained in our experiments are shown in Table 1. We analyze varying two primary dimensions:

The context data extracted from the patient's record and used as input: in addition to T, we study the effect of adding the hint (H), demographic data (D), medications (M), lab-results (L), and previous notes (Np). 2. the model architecture used: Transformer encoder-decoder (T-ED), or with memory compressed attention (T-DMCA).

Every metric for every model wet rained in our experiments is displayed in Table 1.

C. T-ED Experiments

Using the Transformer encoder-decoder architecture, we examine the impact of adding additional context data on performance in the first set of experiments. Without the hint, we find that overall performance is rather low, essentially making this a language model that is unconditional. Thenote's content can be inferred from the first ten sub-word tokens, which also might provide a strong indication of the template that will be used.All things considered, we found that the T-ED Models 3-6 were unable to fully utilize the extra context that was supplied in addition to note-type, hint, and demographic information. The self-reported perplexity of the model is marginally worse, indicating optimization problems. Longer sequence modeling tasks are hard for the T-ED model to handle. Localized attention is used in the T-DMCAarchitecture, which reduces the need for extra attention weights and makes optimization easier in real life.



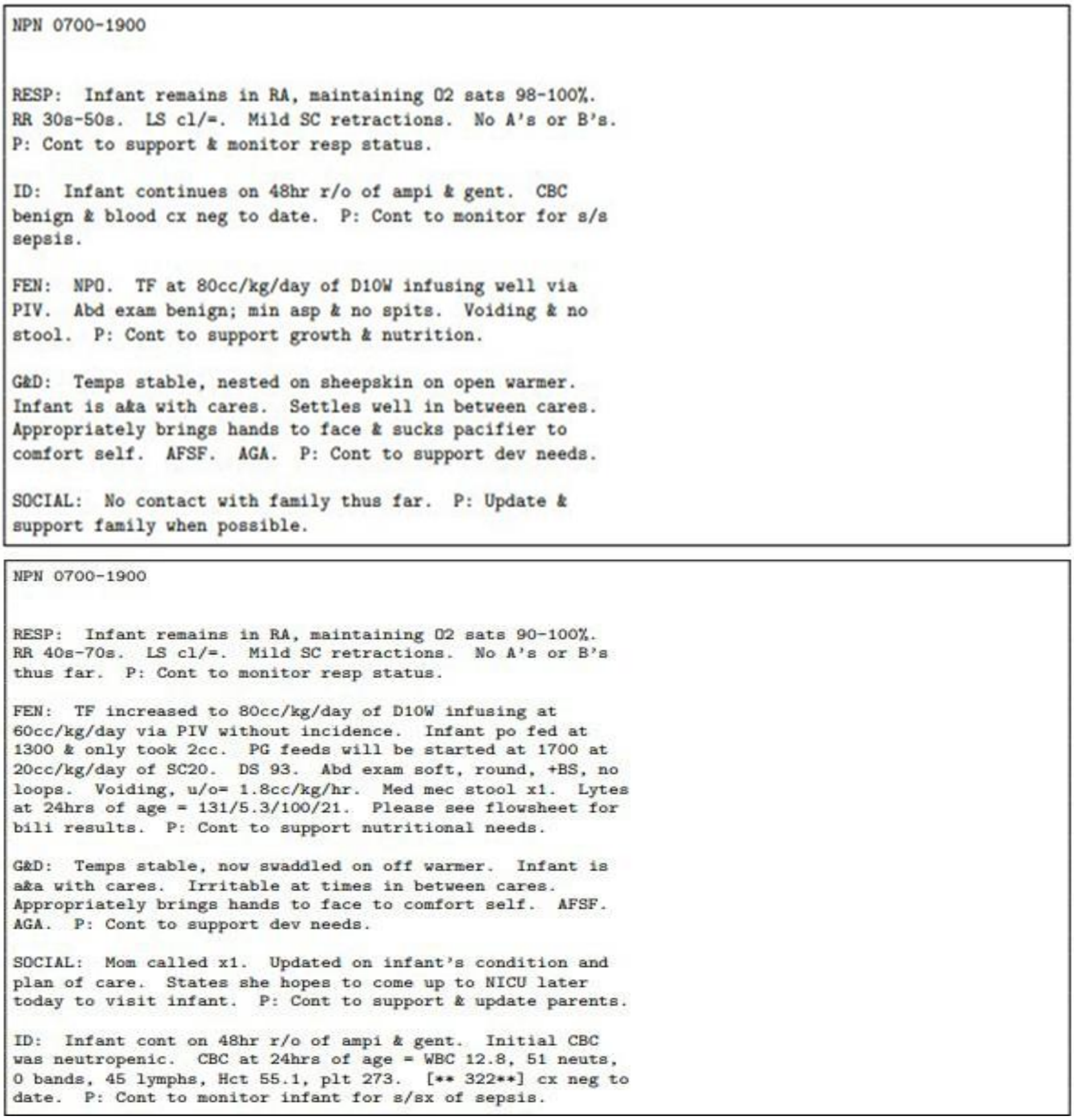
```
[**2101-7-12**] 5:44 PM
 CT HEAD W/O CONTRAST                                          Clip # [**Clip Number (Radiology)
     105293**]
 Reason: eval for ICH
 ______________________________________________________________________________
 [**Hospital 2**] MEDICAL CONDITION:
  History: 79M with fall
 REASON FOR THIS EXAMINATION:
  eval for ICH
 No contraindications for IV contrast
 ______________________________________________________________________________
 WET READ: [**First Name9 (NamePattern2) 1333**] [**Doctor First Name 141**] [**2101-7-12**] 6:02 PM
  no acute intracranial process

 WET READ VERSION #1

 ______________________________________________________________________________
                                 FINAL REPORT
 HISTORY:  79-year-old male with fall.

 COMPARISON:  None available.

 TECHNIQUE:  Non-contrast head CT was obtained.

 FINDINGS:  There is no evidence of acute intracranial hemorrhage, mass effect,
 shift of normally midline structures, hydrocephalus, or acute major vascular
 territorial infarction.  The [**Doctor Last Name 107**]-white matter differentiation is preserved.
 The visualized paranasal sinuses and mastoid air cells are well aerated.

 IMPRESSION:  No acute intracranial process.
```

```
[**2101-7-12**] 5:44 AM
 CHEST (PORTABLE AP)                                           Clip # [**Clip Number (Radiology)
     44638**]
 Reason: ETT,collapsed lung
 Admitting Diagnosis: CHEST PAIN
 ______________________________________________________________________________
 [**Hospital 2**] MEDICAL CONDITION:
  55 year old man s/p mainstem stent placement intubated
 REASON FOR THIS EXAMINATION:
  ETT,collapsed lung
 ______________________________________________________________________________
                                 FINAL REPORT
 INDICATION:  55-year-old man status post mainstem stent placement, ET tube,
 collapsed lung.

 COMPARISON:  Chest radiograph from [**2101-7-11**].

 FINDINGS:  ET tube is 8.2 cm above the carina.  A left mainstem bronchus stent
 is in place.  Since the prior radiograph, there is no significant change.
 Small left pleural effusion is unchanged The right lung is clear.  There is no
 focal consolidation, or pneumothorax.  The bony structures are intact.

 IMPRESSION:  ET tube 8.2 cm above the carina.  Otherwise, no significant
 change since prior radiograph.
```

Fig. 3: A radiology note conditioned on note-type and hint, produced by Model 2(fromTable1). The accurate global structure of these notes are captured by the model

D. T-DMCA Experiments

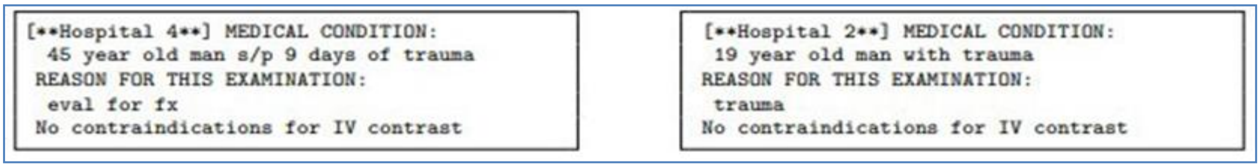
Results for the T-DMCA model in the longest note-context cases are displayed in Models 7-9. The outcomes surpass T-ED, and the optimal model is the one with the greatest amount of context, demonstrating that, in contrast to the T-ED model, the architecture can suitably benefit from more data. Overall, the most contextualized T-DMCAoutperforms all others in terms of metrics. Due to their local nature, the perplexity and next-token accuracy metrics demonstrate a smaller relative improvement; predicting the very next token does not frequently require long-range dependencies. However, because of the compounding of errors when predicting the entire note, the ROUGE-based scores demonstrate a larger relative improvement. A complete example nursing progress note about a newborn patient created by the best Model 9 is displayed in Figure 4. With the same section headings as the ground truth note, it has deduced the correct global structure of these notes. There is a lot of vocabulary overlap and good emulation of style within each section.



```
NPN 0700-1900


RESP:  Infant remains in RA, maintaining O2 sats 98-100%.
RR 30s-50s.  LS cl/=.  Mild SC retractions.  No A's or B's.
P: Cont to support & monitor resp status.

ID:  Infant continues on 48hr r/o of ampi & gent.  CBC
benign & blood cx neg to date.  P: Cont to monitor for s/s
sepsis.

FEN:  NPO.  TF at 80cc/kg/day of D10W infusing well via
PIV.  Abd exam benign; min asp & no spits.  Voiding & no
stool.  P: Cont to support growth & nutrition.

G&D:  Temps stable, nested on sheepskin on open warmer.
Infant is a&a with cares.  Settles well in between cares.
Appropriately brings hands to face & sucks pacifier to
comfort self.  AFSF.  AGA.  P: Cont to support dev needs.

SOCIAL:  No contact with family thus far.  P: Update &
support family when possible.
```

```
NPN 0700-1900


RESP:  Infant remains in RA, maintaining O2 sats 90-100%.
RR 40s-70s.  LS cl/=.  Mild SC retractions.  No A's or B's
thus far.  P: Cont to monitor resp status.

FEN:  TF increased to 80cc/kg/day of D10W infusing at
60cc/kg/day via PIV without incidence.  Infant po fed at
1300 & only took 2cc.  PG feeds will be started at 1700 at
20cc/kg/day of SC20.  DS 93.  Abd exam soft, round, +BS, no
loops.  Voiding, u/o= 1.8cc/kg/hr.  Med mec stool x1.  Lytes
at 24hrs of age = 131/5.3/100/21.  Please see flowsheet for
bili results.  P: Cont to support nutritional needs.

G&D:  Temps stable, now swaddled on off warmer.  Infant is
a&a with cares.  Irritable at times in between cares.
Appropriately brings hands to face to comfort self.  AFSF.
AGA.  P: Cont to support dev needs.

SOCIAL:  Mom called x1.  Updated on infant's condition and
plan of care.  States she hopes to come up to NICU later
today to visit infant.  P: Cont to support & update parents.

ID:  Infant cont on 48hr r/o of ampi & gent.  Initial CBC
was neutropenic.  CBC at 24hrs of age = WBC 12.8, 51 neuts,
0 bands, 45 lymphs, Hct 55.1, plt 273.  [** 322**] cx neg to
date.  P: Cont to monitor infant for s/sx of sepsis.
```

Fig. 4: (Nursing) note was generated from Model 9 (fromTable 1)



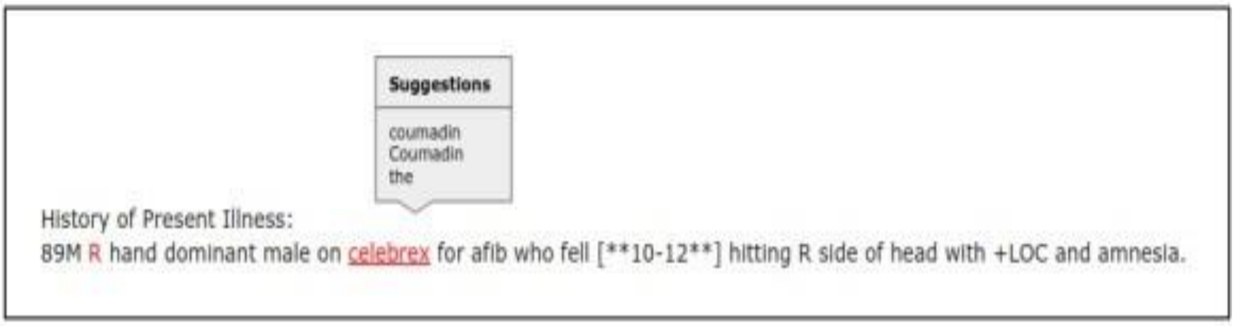
```
[**Hospital 4**] MEDICAL CONDITION:
  45 year old man s/p 9 days of trauma
REASON FOR THIS EXAMINATION:
  eval for fx
No contraindications for IV contrast
```

```
[**Hospital 2**] MEDICAL CONDITION:
  19 year old man with trauma
REASON FOR THIS EXAMINATION:
  trauma
No contraindications for IV contrast
```

Fig. 5: (Left) Segment of a note describingage and gender of a patient

IV. DETECTION OF ERRORS IN NOTES

As explained previously, our models that are conditioned on demographic data can produce theright age given the structured data that is fed into the model. The identical section from the generated and actual notes, for instance, is shown in Figure 5, where only the generated note is accurate.We discuss how the model might be used to identify and stop note errors, as this could be viewedasamistakeintheoriginalnote.Wetamperedwithafewnotesbysubstitutingrandomdrug names for specific ones in order to mimic errors and show how error detection can be applied to preexisting notes. A sample of note segments from the test set is displayed in Figure 6, with low likelihood words indicated in red. We show you the best replacement options for the words we tainted. Usually, just the tainted word is indicated, and suggestions either contain the original word or synonyms.

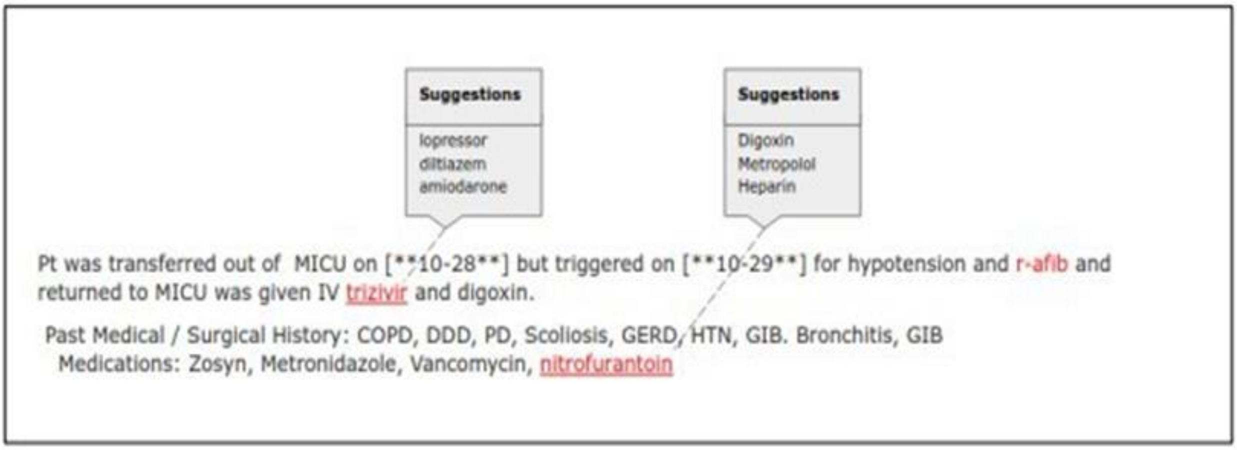


Suggestions

coumadin
Coumadin
the

History of Present Illness:
89M R hand dominant male on celebrex for afib who fell [**10-12**] hitting R side of head with +LOC and amnesia.

- Coumarin was replaced with celebrex, and the model suggests coumadin, which is an alternate name



Suggestions

lopressor
diltiazem
amiodarone

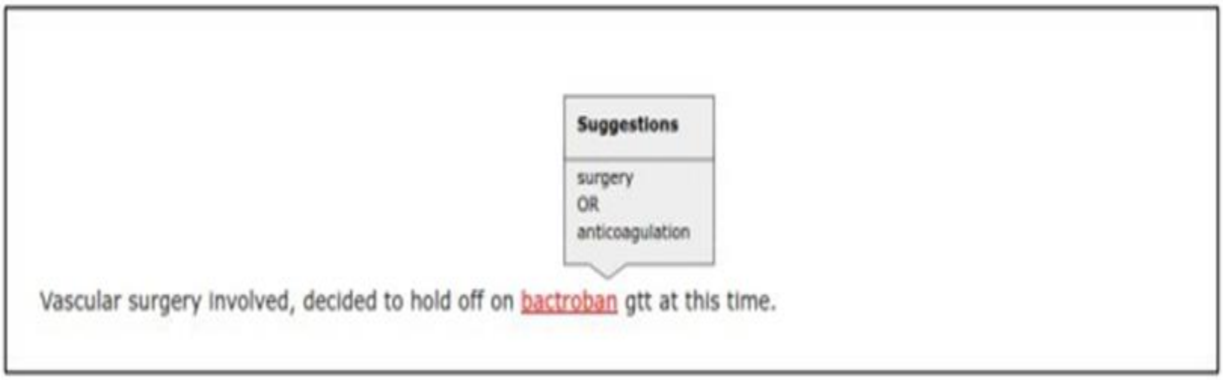
Suggestions

Digoxin
Metropolol
Heparin

Pt was transferred out of MICU on [**10-28**] but triggered on [**10-29**] for hypotension and r-afib and returned to MICU was given IV trizivir and digoxin.

Past Medical / Surgical History: COPD, DDD, PD, Scoliosis, GERD, HTN, GIB. Bronchitis, GIB
Medications: Zosyn, Metronidazole, Vancomycin, nitrofurantoin

- Amiodarone is one of correct suggestions for trizivir, while suggestions for nitrofurantoin arerelated heart medications to the original, lasix.



Suggestions

surgery
OR
anticoagulation

Vascular surgery involved, decided to hold off on bactroban gtt at this time.

Fig. 6: Three corrupted snippets each from a different test note from MIMIC-III. Words highlighted in red are deemed unlikely under the model conditioned on medications.

- The original drug was heparin, an anti coagulation drug

Underlined words were the result of artificial corruption and the suggested model corrections are shown.

V. LIMITATIONS AND FUTURE WORK

Although it's possible that more prior context would be instructive, in our experiments we limited our context data to the 24 hours prior to the note in order to limit input sequence length for performance reasons. Many events may happen in a 24-hour period because of the intensive nature of the ICU; however, a larger window of context should be used for datasets that are not from the ICU. Additionally, although this was not investigated, more columns from additional tables in MIMIC-III could be added as note-context. Frequently, the EHR's maximum context is insufficient to completely predict the note. The most notable example is the absence of imaging data for radiology reports in MIMIC-III. We also do not have up-to-date information on patient-provider interactions for non-imaging notes. Subsequent investigations may aim to supplement the note-context with extraneous data sources, such as imaging records or transcripts of patient-physician exchanges. Future work could attempt to augment the note-context with data beyond the EHR, e.g. imaging data, or transcripts of patient-doctor interactions.

VI. CONCLUSION

We have demonstrated how to represent the multi-modal data context to the model and presented a new language modeling task for clinical notes based on HER data. We provided encouraging results demonstrating the predictive capacity of such models and suggested evaluation metrics for the task. We talked abouthow these models could be helpful for advanced auto-complete and spell-checking functions, which could lessen the workload associated with creating accurate clinicaldocumentation.

REFERENCES

[1]. Christine Sinsky, Lacey Colligan, Ling Li, MirelaPrgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Annals of internal medicine, 165(11):753–760, 2016.

[2]. George A Gellert, Ricardo Ramirez, and S Luke Webster. The rise of the medical scribe industry: implications for the advancement of electronic health records. Jama, 313(13): 1315–1316, 2015.

[3]. Mohammed Saeed, Mauricio Villarroel,Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. Critical care medicine, 39(5):952, 2011.

[4]. Tal Mandelbaum, Daniel J Scott, Joon Lee, Roger G Mark, Atul Malhotra, Sushrut S Waikar, Michael D Howell, and Daniel Talmor. Outcome of critically ill patients with acute kidney injury using the akin criteria. Critical care medicine, 39(12):2659, 2011.

[5]. Alistair EW Johnson,TomJPollard,andRogerGMark.Reproducibilityincriticalcare:amortality prediction case study. In Machine Learning for Healthcare Conference, pages 361–376, 2017.

[6]. Edward Choi, Mohammad TahaBahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference, pages 301–318, 2016.

[7]. Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics, 2017.

[8]. Fran¸coisPortet, Ehud Reiter, Albert Gatt, Jim Hunter, SomayajuluSripada, Yvonne Freer, and Cindy Sykes. Automatic generation of textual summaries from neonatal intensive care data. Artificial Intelligence, 173(7-8):789–816, 2009.

[9]. Baoyu Jing, PengtaoXie, and Eric Xing. On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195, 2017.

[10]. Alvin Rajkomar, Eyal Oren, Kai Chen,Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1):18, 2018

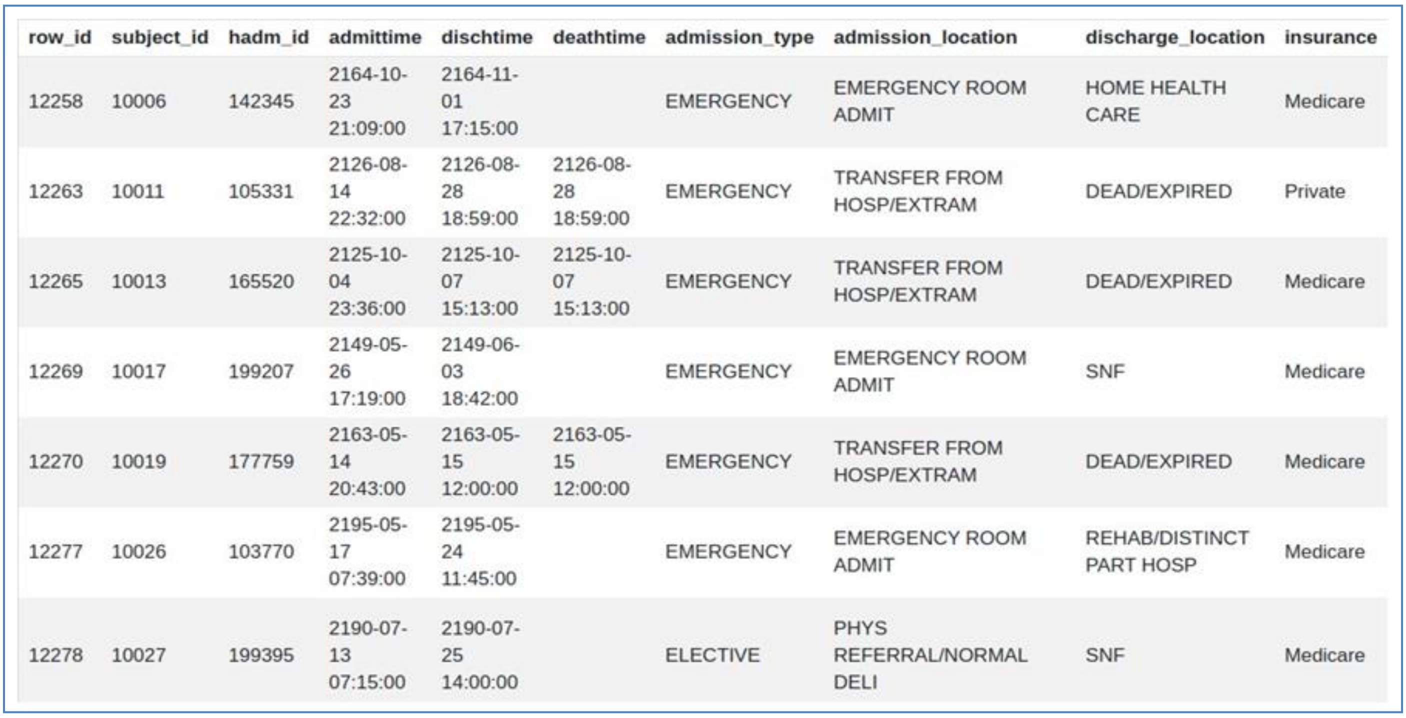
[11]. YonghuiWu,MikeSchuster,ZhifengChen,QuocVLe,MohammadNorouzi,WolfgangMacherey, Maxim Krikun, Yuan Cao, Qin Gao, KlausMacherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.

[12]. Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, RohitPrabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Katya Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. arXiv preprint arXiv:1712.01769, 2017

[13]. Johnson, A., Pollard, T., and Mark, R. (2016) 'MIMIC-III Clinical Database' (version 1.4), PhysioNet. Available at: https://doi.org/10.13026/C2XW26.

APPENDIX

A. Sample table from Mimic III dataset



row_id	subject_id	hadm_id	admittime	dischtime	deathtime	admission_type	admission_location	discharge_location	insurance
12258	10006	142345	2164-10-23 21:09:00	2164-11-01 17:15:00		EMERGENCY	EMERGENCY ROOM ADMIT	HOME HEALTH CARE	Medicare
12263	10011	105331	2126-08-14 22:32:00	2126-08-28 18:59:00	2126-08-28 18:59:00	EMERGENCY	TRANSFER FROM HOSP/EXTRAM	DEAD/EXPIRED	Private
12265	10013	165520	2125-10-04 23:36:00	2125-10-07 15:13:00	2125-10-07 15:13:00	EMERGENCY	TRANSFER FROM HOSP/EXTRAM	DEAD/EXPIRED	Medicare
12269	10017	199207	2149-05-26 17:19:00	2149-06-03 18:42:00		EMERGENCY	EMERGENCY ROOM ADMIT	SNF	Medicare
12270	10019	177759	2163-05-14 20:43:00	2163-05-15 12:00:00	2163-05-15 12:00:00	EMERGENCY	TRANSFER FROM HOSP/EXTRAM	DEAD/EXPIRED	Medicare
12277	10026	103770	2195-05-17 07:39:00	2195-05-24 11:45:00		EMERGENCY	EMERGENCY ROOM ADMIT	REHAB/DISTINCT PART HOSP	Medicare
12278	10027	199395	2190-07-13 07:15:00	2190-07-25 14:00:00		ELECTIVE	PHYS REFERRAL/NORMAL DELI	SNF	Medicare

Fig. 6: Sample Table from the Mimic III dataset

B. A2.MIMIC-IIITablesandcolumnsused

- **Patients:** GENDER(sex), DOB(dateof birth)
- **Prescriptions:** DRUG (drug name)
- **Note Events:** CATEGORY,TEXT (raw text of note)
- **Lab Events:** ITEMID(join key to DLABITEMS), VALUE, VALUEUOM, FLAG
- **DLABITEMS:** LABEL (labname)