DrugEHRQA: A Question Answering Dataset on Structured and Unstructured Electronic Health Records For Medicine Related Queries

-Jayetri Bardhan, Anthony Colas, Kirk Roberts, Daisy Zhe Wang

Introduction

- An Electronic Health Record (EHR) is an electronic version of a patient's medical history.
- They contain information about the patients' demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports.
- EHR data can be stored in the form of structured data (KGs, multi-relational tables) or as unstructured clinical notes.
- The information in structured and unstructured EHRs is not strictly disjoint: information may be -
    - duplicated
    - contradictory, or
    - provide additional context between these sources.

Proposed dataset

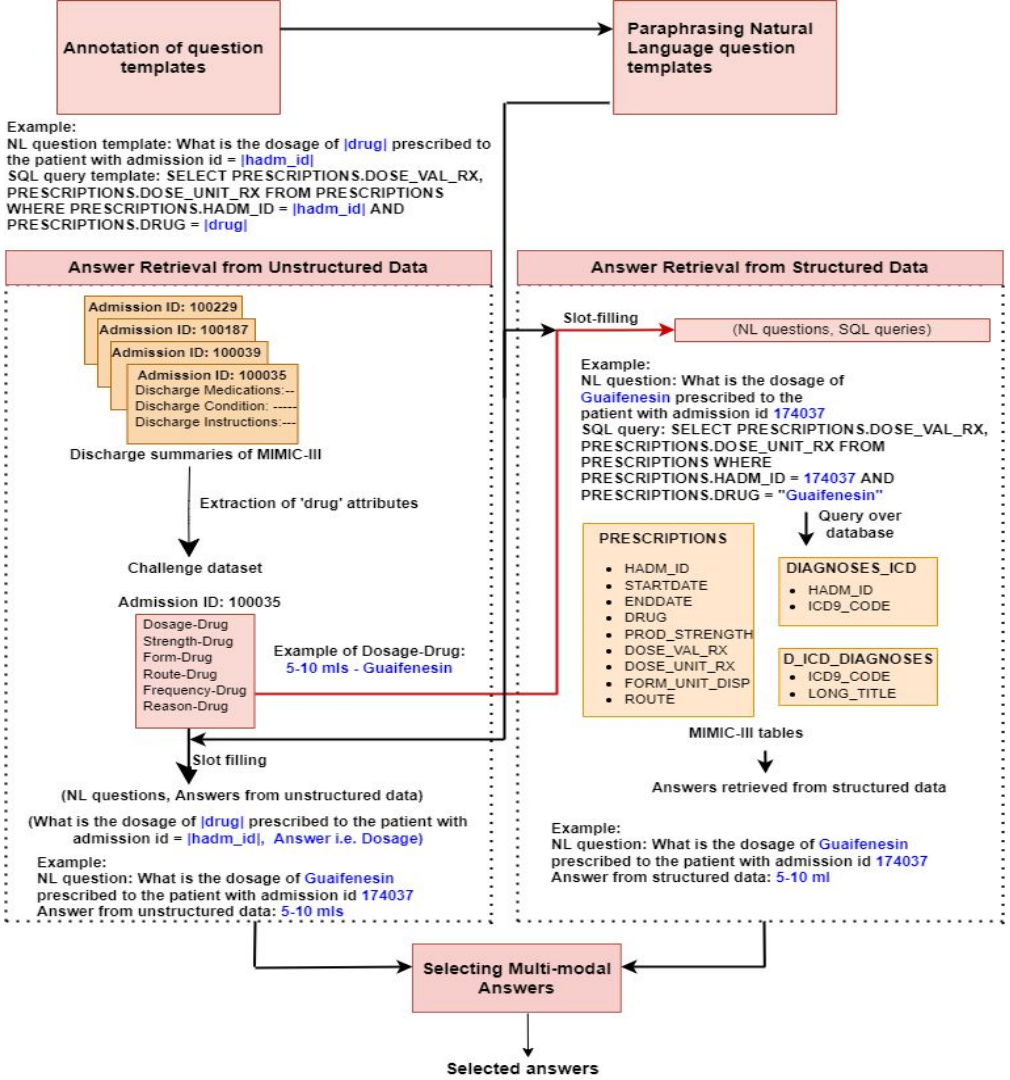
- Developed the first question answering dataset (DrugEHRQA) containing question answer pairs from both structured tables and unstructured notes from a publicly available EHR database, MIMIC-III.
- Dataset contains:
    - Natural language (NL) questions,
    - SQL queries
    - Retrieved Answer from one or both modalities
    - Selected multimodal answer
- Novel technique to generate multimodal QA dataset using existing annotations of a non-QA application.

Previous datasets for QA on EHRs

- QA on knowledge bases (KBs)
  - ClinicalKBQA (Wang et al.,2021)
- QA on EHR tables
  - MIMICSQL (Wang et al., 2020b) , emrKBQA (Raghavan et al., 2021)
- QA on clinical notes
  - emrQA (Pampari et al., 2018), CliniQG4QA (Yue et al., 2021)

Lack of any existing multimodal QA dataset on EHRs!

Dataset generation framework



NL Question templates derived from drug-related entities and attributes extracted from the clinical notes using the challenge dataset

Drug attributes and entities	Examples	NL Question templates
Drug	Lithium Carbonate, Propafenone	What are the list of medicines prescribed to the patient
Strength-Drug	(300mg, Lithium Carbonate)	What is the drug strength of |drug|
Form-Drug	(Tablet, Propafenone)	What is the form of |drug|
Route-Drug	(PO, Metoprolol Tartrate)	What is the route of administration for the drug |drug|
Dosage-Drug	(One tablet, Bactrim)	What is the dosage of |drug| prescribed to the patient
Frequency-Drug	(14 day, Zosyn)	How long has the patient been taking |drug|
Reason-Drug	(Constipation, Polyethylene Glycol)	Why is the patient been given |drug|
Reason-Drug	(Polyethylene Glycol, Constipation)	What is the medication prescribed to the patient for |problem|
Reason-Drug, Dosage-Drug	(Constipation, Polyethylene Glycol), (300mg , Polyethylene Glycol)	List all the medicines and their dosages prescribed to the patient for |problem|

Paraphrasing Natural Language Questions

What is the medication prescribed to the patient with admission id |hadm_id| for |problem|
Which medicines are taken by the patient suffering from |problem| having an admission id of |hadm_id|
For |problem|, name the drugs that has been recommended to be taken by the patient with admission id = |hadm_id|
What medication is the patient with an admission id of |hadm_id| taking for |problem|

Rules for selecting multi-modal answers

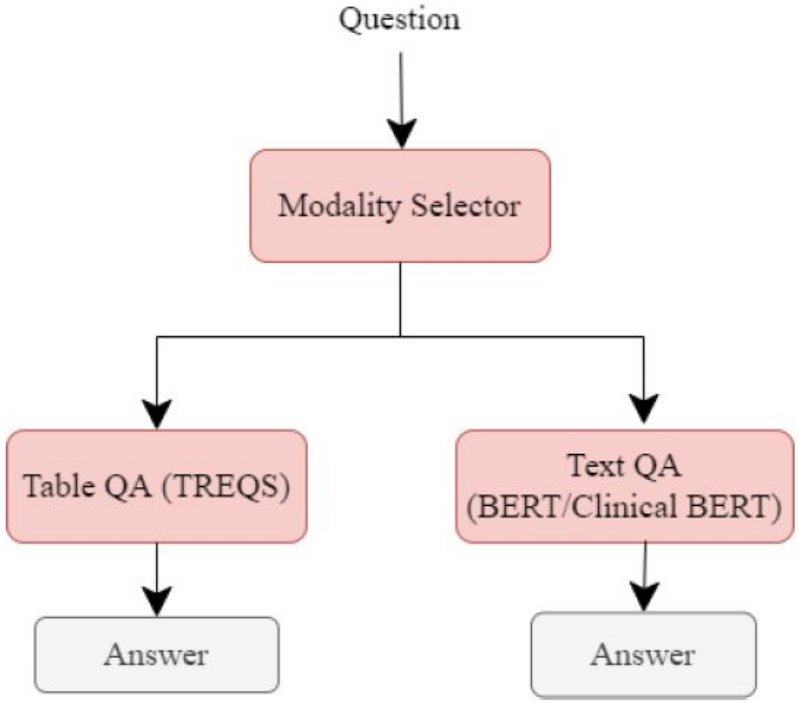
- If the answer exists in only one modality, the available answer is selected as the multi-modal answer.
- Check for overlapping answers.
    - If there is even one common answer between "Answer Structured" and "Answer Unstructured", choose the common answer.
- If there are no common answers between the two modalities, choose the answer from the modality which is more reliable.

Applying rules for selecting multi-modal answers

Question	Answer from Structured	Answerfrom Unstructured	Multi-modal answer
WHAT IS THE MEDICATION PRESCRIBED TO THE PATIENT WITH ADMISSION ID 111160 FOR PAIN	–	MORPHINE	MORPHINE
WHAT IS THE DRUG STRENGTH OF SIMETHICONE PRESCRIBED TO THE PATIENT WITH ADMISSION ID 125206	80MG TABLET	80 MG	80MG TABLET
HOW LONG HAS THE PATIENT WITH ADMISSION ID = 187782 BEEN TAKING VANCOMYCIN	14 DAYS	14 DAYS	14 DAYS
WHAT IS THE DRUG STRENGTH OF FUROSEMIDE PRESCRIBED TO THE PATIENT WITH ADMISSION ID 100509	40MG/4ML VIAL	10 MG	40MG/4ML VIAL

Proposed baseline model (MultimodalEHRQA)

- (MultimodalEHRQA) uses the predictions of a modality selection network to choose between EHR tables and clinical notes to answer the questions.
- This is used to direct the questions to the table-based or text-based state-of-the-art QA model.



Multimodal Selection Network

- The multimodal selection network uses a binary classification approach
- BERT with a feedforward network followed by a softmax layer is used to predict the correct or the more reliable modality.

QA models

- **TREQS:**
  - TRanslate-Edit Model for Question-to-SQL (Wang et al., 2020b) is a sequence-to-sequence model which generates SQL query for a given question.
- **RAT-SQL:**
  - Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers (Wang et al., 2020a) was used in order to address the more complex, nested SQL queries of the DrugEHRQA dataset.
- **BERT QA (Devlin et al., 2019) and ClinicalBERT QA** (Alsentzer et al., 2019) are used for QA over unstructured EHR data.

Results of multimodal QA

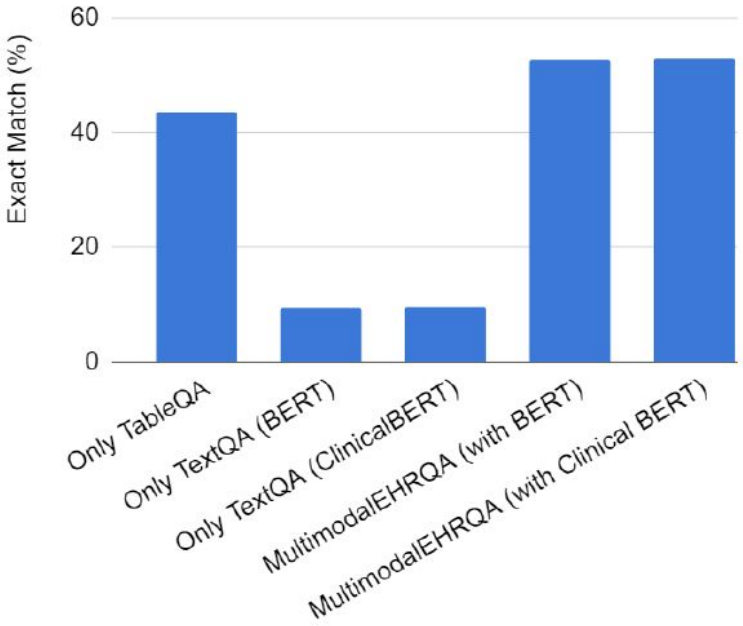


Figure (a): Exact match values of MultimodalEHRQA in comparison to single-modal QA models for questions with non-overlapping answers.

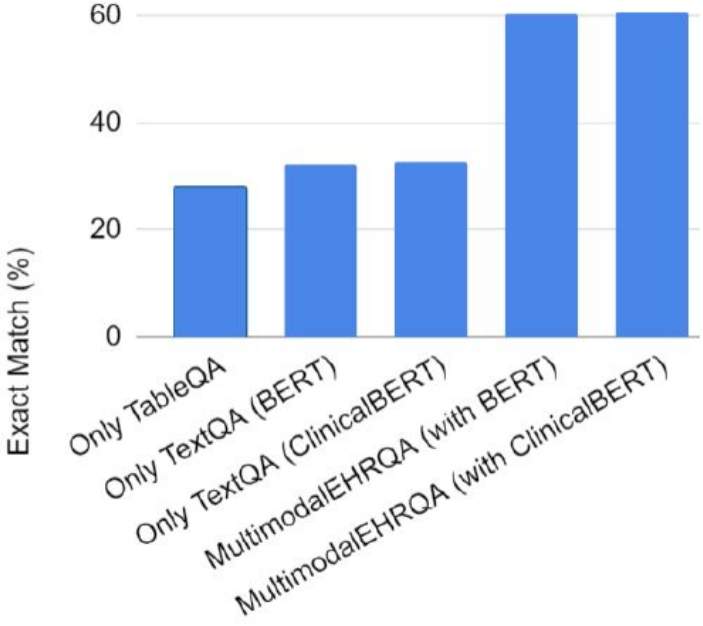


Figure (b): Overall performance of MultimodalEHRQA in comparison to single-modal QA models for the entire DrugEHRQA dataset (basic version).

Limitations

- The dataset generation technique is limited only to the MIMIC-III database. The same steps cannot be reproduced for other EHR databases.
- Diversity of the questions in the DrugEHRQA dataset are limited by the type of relations extracted from the challenge dataset.

Conclusion and Future Work

- The DrugEHRQA dataset introduces new horizons of research in multimodal QA over EHRs.
- Introduced a simple baseline model for multimodal QA on EHRs.
- In the future, we will try to work on multimodal QA models for EHRs which jointly trains the model on both table and text.