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

**Annotation of question templates**

Example:

NL question template: What is the dosage of [drug] prescribed to the patient with admission id = [hadm_id].

SQL query template: SELECT PRESCRIPTIONS.DOSE_VAL_RX, PRESCRIPTIONS.DOSE_UNIT_RX FROM PRESCRIPTIONS WHERE PRESCRIPTIONS.HADM_ID = [hadm_id] AND PRESCRIPTIONS.DRUG = [drug].

**Paraphrasing Natural Language question templates**

**Answer Retrieval from Unstructured Data**

- **Discharge summaries of MIMIC-III**
- **Challenge dataset**
- **Example of Dosage-Drug:**
  - 5-10 ml - Guaiifenesin

**Slot filling**

(NL questions, Answers from unstructured data)

(What is the dosage of [drug]) prescribed to the patient with admission id = [hadm_id], Answer i.e. Dosage)

**Selecting Multi-modal Answers**

Selected answers

**Answer Retrieval from Structured Data**

- **Example:**
  - NL question: What is the dosage of Guaiifenesin prescribed to the patient with admission id 174037
  - Answer from structured data: 5-10 ml

**Slot filling**

(NL questions, SQL queries)

**Query over database**

- **MIMIC-III tables**
- **Example:**
  - NL question: What is the dosage of Guaiifenesin prescribed to the patient with admission id 174037
  - Answer from structured data: 5-10 ml
NL Question templates derived from drug-related entities and attributes extracted from the clinical notes using the challenge dataset

<table>
<thead>
<tr>
<th>Drug attributes and entities</th>
<th>Examples</th>
<th>NL Question templates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>Lithium Carbonate, Propafenone (300mg, Lithium Carbonate)</td>
<td>What are the list of medicines prescribed to the patient</td>
</tr>
<tr>
<td>Strength-Drug</td>
<td>Tablet, Propafenone</td>
<td>What is the drug strength of ldrug</td>
</tr>
<tr>
<td>Form-Drug</td>
<td>PO, Metoprolol Tartrate</td>
<td>What is the form of ldrug</td>
</tr>
<tr>
<td>Route-Drug</td>
<td>One tablet, Bactrim</td>
<td>What is the route of administration for the drug ldrug</td>
</tr>
<tr>
<td>Dosage-Drug</td>
<td>14 day, Zosyn</td>
<td>What is the dosage of ldrug</td>
</tr>
<tr>
<td>Frequency-Drug</td>
<td>Constipation, Polyethylene Glycol</td>
<td>How long has the patient been taking ldrug</td>
</tr>
<tr>
<td>Reason-Drug</td>
<td>Polyethylene Glycol, Constipation</td>
<td>Why is the patient been given ldrug</td>
</tr>
<tr>
<td>Reason-Drug</td>
<td>Constipation, Polyethylene Glycol</td>
<td>What is the medication prescribed to the patient for lproblem</td>
</tr>
<tr>
<td>Reason-Drug, Dosage-Drug</td>
<td>Polyethylene Glycol, (300mg, Polyethylene Glycol)</td>
<td>List all the medicines and their dosages prescribed to the patient for lproblem</td>
</tr>
</tbody>
</table>
Paraphrasing Natural Language Questions

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the medication prescribed to the patient with admission id</td>
</tr>
<tr>
<td>Which medicines are taken by the patient suffering from</td>
</tr>
<tr>
<td>For</td>
</tr>
<tr>
<td>What medication is the patient with an admission id of</td>
</tr>
</tbody>
</table>
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

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