emrQA: A Large Corpus for Question Answering on Electronic Medical Records

Anusri Pampari[*] **Preethi Raghavan**[†♣] **Jennifer Liang**[†] **and Jian Peng**[*‡]
[♣]MIT-IBM Watson AI Lab, Cambridge, MA
[†]IBM TJ Watson Research Center, Yorktown Heights, NY
[*]Dept. of Computer Science, University of Illinois Urbana Champaign, IL
[‡]Carle Illinois College of Medicine, University of Illinois Urbana Champaign, IL
[*]{pampari2,jianpeng}@illinois.edu [†]{praghav,jjliang}@us.ibm.com

Abstract

We propose a novel methodology to generate domain-specific large-scale question answering (QA) datasets by re-purposing existing annotations for other NLP tasks. We demonstrate an instance of this methodology in generating a large-scale QA dataset for electronic medical records by leveraging existing expert annotations on clinical notes for various NLP tasks from the community shared i2b2 datasets[§]. The resulting corpus (emrQA) has 1 million questions-logical form and 400,000+ question-answer evidence pairs. We characterize the dataset and explore its learning potential by training baseline models for question to logical form and question to answer mapping.

1 Introduction

Automatic question answering (QA) has made big strides with several open-domain and machine comprehension systems built using large-scale annotated datasets (Voorhees et al., 1999; Ferrucci et al., 2010; Rajpurkar et al., 2016; Joshi et al., 2017). However, in the clinical domain this problem remains relatively unexplored. Physicians frequently seek answers to questions from unstructured electronic medical records (EMRs) to support clinical decision-making (Demner-Fushman et al., 2009). But in a significant majority of cases, they are unable to unearth the information they want from EMRs (Tang et al., 1994). Moreover to date, there is no general system for answering natural language questions asked by physicians on a patient's EMR (Figure 1) due to lack of large-scale datasets (Raghavan and Patwardhan, 2016).

EMRs are a longitudinal record of a patient's health information in the form of unstructured clinical notes (progress notes, discharge summaries etc.) and structured vocabularies. Physicians wish to answer questions about medical entities and relations from the EMR, requiring a deeper understanding of clinical notes. While this may be likened to machine comprehension, the longitudinal nature of clinical discourse, little to no redundancy in facts, abundant use of domain-specific terminology, temporal narratives with multiple related diseases, symptoms, medications that go back and forth in time, and misspellings, make it complex and difficult to apply existing NLP tools (Demner-Fushman et al., 2009; Raghavan and Patwardhan, 2016). Moreover, answers may be implicit or explicit and may require domain-knowledge and reasoning across clinical notes. Thus, building a credible QA system for patient-specific EMR QA requires large-scale question and answer annotations that sufficiently capture the challenging nature of clinical narratives in the EMR. However, serious privacy concerns about sharing personal health information (Devereaux, 2013; Krumholz et al., 2016), and the tedious nature of assimilating answer annotations from across longitudinal clinical notes, makes this task impractical and possibly erroneous to do manually (Lee et al., 2017).

Record Date: 08/09/98

08/31/96 ascending aortic root replacement with homograft with omentopexy. The patient continued to be hemodynamically stable making good progress. Physical examination: BMI: 33.4 Obese, high risk. Pulse: 60. resp. rate: 18

Question: Has the patient ever had an abnormal BMI?
Answer: BMI: 33.4 Obese, high risk
Question: When did the patient last receive a homograft replacement ?
Answer: 08/31/96 ascending aortic root replacement with homograft with omentopexy.

Figure 1: Question-Answer pairs from emrQA clinical note.

In this work, we address the lack of any publicly available EMR QA corpus by creating a large-scale dataset, emrQA, using a novel gener-

[*]This work was conducted during an internship at IBM
[§]https://www.i2b2.org/NLP/DataSets/

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2357–2368
Brussels, Belgium, October 31 - November 4, 2018.

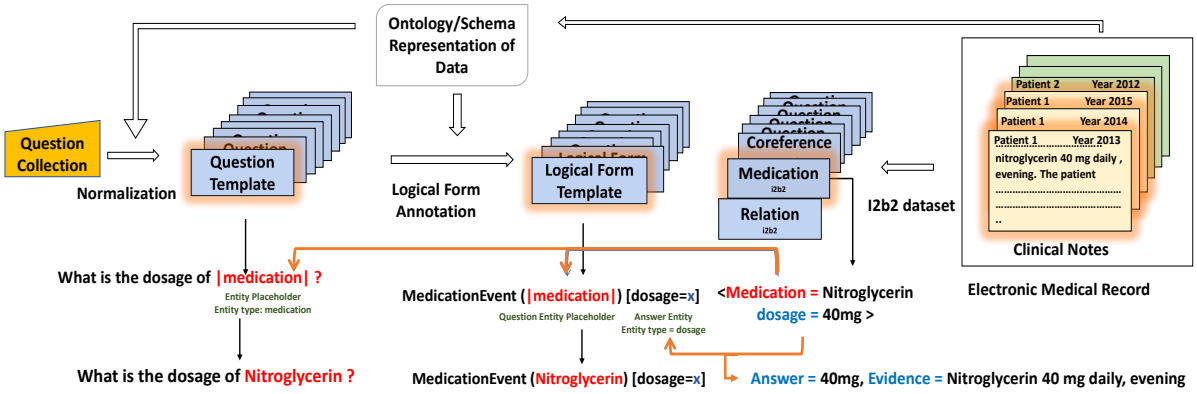

Figure 2: Our QA dataset generation framework using existing i2b2 annotations on a given patient's record to generate a question, its logical form and answer evidence. The highlights in the figure show the annotations being used for this example.

ation framework that allows for minimal expert involvement and re-purposes existing annotations available for other clinical NLP tasks (i2b2 challenge datasets (Guo et al., 2006)). The annotations serve as a proxy-expert in generating questions, answers, and logical forms. Logical forms provide a human-comprehensible symbolic representation, linking questions to answers, and help build interpretable models, critical to the medical domain (Davis et al., 1977; Vellido et al., 2012). We analyze the emrQA dataset in terms of question complexity, relations, and the reasoning required to answer questions, and provide neural and heuristic baselines for learning to predict question-logical forms and question-answers.

The main contributions of this work are as follows:

- A novel framework for systematic generation of domain-specific large-scale QA datasets that can be used in any domain where manual annotations are challenging to obtain but limited annotations may be available for other NLP tasks.
- The first accessible patient-specific EMR QA dataset, emrQA*, consisting of 400,000 question-answer pairs and 1 million question-logical form pairs. The logical forms will allow users to train and benchmark interpretable models that justify answers with corresponding logical forms.
- Two new reasoning challenges, namely arithmetic and temporal reasoning, that are absent in open-domain datasets like SQuAD (Rajpurkar et al., 2016).

*https://github.com/panushri25/emrQA, scripts to generate emrQA from i2b2 data. i2b2 data is accessible by everyone subject to a license agreement.

2 Related Work

Question Answering (QA) datasets are classified into two main categories: (1) machine comprehension (MC) using unstructured documents, and (2) QA using Knowledge Bases (KBs).

MC systems aim to answer any question that could be posed against a reference text. Recent advances in crowd-sourcing and search engines have resulted in an explosion of large-scale (100K) MC datasets for factoid QA, having ample redundant evidence in text (Rajpurkar et al., 2016; Trischler et al., 2016; Joshi et al., 2017; Dhingra et al., 2017). On the other hand, complex domain-specific MC datasets such as MCTest (Richardson et al., 2013), biological process modeling (Berant et al., 2014), BioASQ (Tsatsaronis et al., 2015), InsuranceQA (Feng et al., 2015), etc have been limited in scale (500-10K) because of the complexity of the task or the need for expert annotations that cannot be crowd-sourced or gathered from the web. In contrast to the open-domain, EMR data cannot be released publicly due to privacy concerns (Šuster et al., 2017). Also, annotating unstructured EMRs requires a medical expert who can understand and interpret clinical text. Thus, very few datasets like i2b2, MIMIC (Johnson et al., 2016) (developed over several years in collaboration with large medical groups and hospitals), share small-scale annotated clinical notes. In this work, we take advantage of the limited expertly annotated resources to generate emrQA.

KB-based QA datasets, used for semantic parsing, are traditionally limited by the requirement of annotated question and logical form (LF) pairs for supervision where the LF are used to retrieve answers from a schema (Cai and Yates, 2013; Lopez et al., 2013; Bordes et al., 2015). Roberts and Demner-Fushman (2016) generated a corpus by

Datasets	#QA	#QL	#notes	Property	Stats.
Relations	141,243	1,061,710	425	Question len.	8.6
Medications	255,908	198,739	261	Evidence len.	18.7
Heart disease	30,731	36,746	119	LF len.	33
Obesity	23,437	280	1,118	Note len.	3825
Smoking	4,518	6	502	# of evidence	1.5
emrQA	455,837	1,295,814	2,425	# Ques. in note	187

Table 1: (left) i2b2 dataset distribution in emrQA, and (right) emrQA properties with length in tokens, averaged

How was the $\|problem\|$ managed ?
How was the patient's $\|problem\|$ treated ?
What was done to correct the patient's $\|problem\|$?
Has the patient ever been treated for a $\|problem\|$?
What treatment has the patient had for his $\|problem\|$?
Has the patient ever received treatment for $\|problem\|$?
What treatments for $\|problem\|$ has this patient tried ?

Table 2: Paraphrase templates of a question type in emrQA.

manually annotating LFs on 468 EMR questions (not released publicly), thus limiting its ability to create large scale datasets. In contrast, we only collect LFs for question templates from a domain-expert - the rest of our corpus is automatically generated.

Recent advances in QA combine logic-based and neural MC approaches to build hybrid models (Usbeck et al., 2015; Feng et al., 2016; Palangi et al., 2018). These models are driven to combine the accuracy of neural approaches (Hermann et al., 2015) and the interpretability of the symbolic representations in logic-based methods (Gao et al.; Chabierski et al., 2017). Building interpretable yet accurate models is extremely important in the medical domain (Shickel et al., 2017). We generate large-scale ground truth annotations (questions, logical forms, and answers) that can provide supervision to learn such hybrid models. Our approach to generating emrQA is in the same spirit as Su et al. (2016), who generate graph queries (logical forms) from a structured KB and use them to collect answers. In contrast, our framework can be applied to generate QA dataset in any domain with minimal expert input using annotations from other NLP tasks.

3 QA Dataset Generation Framework

Our general framework for generating a large-scale QA corpus given certain resources consists of three steps: (1) collecting questions to capture domain-specific user needs, followed by normalizing the collected questions to templates by replacing entities (that may be related via binary or composite relations) in the question with placeholders. The entity types replaced in the question are grounded in an ontology like WordNet (Miller, 1995), UMLS (Bodenreider, 2004), or a user-generated schema that defines and relates different entity types. (2) We associate question templates with expert-annotated logical form templates; logical forms are symbolic representations using relations from the ontology/schema to express the relations in the question, and associate the question entity type with an answer entity type. (3) We then proceed to the important step of re-purposing existing NLP annotations to populate question-logical form templates and generate answers. QA is a complex task that requires addressing several fundamental NLP problems before accurately answering a question. Hence, obtaining expert manual annotations in complex domains is infeasible as it is tedious to expert-annotate answers that may be found across long document collections (e.g., longitudinal EMR) (Lee et al., 2017). Thus, we reverse engineer the process where we reuse expert annotations available in NLP tasks such as entity recognition, coreference, and relation learning, based on the information captured in the logical forms to populate entity placeholders in templates and generate answers. Reverse engineering serves as a proxy expert ensuring that the generated QA annotations are credible. The only manual effort is in annotating logical forms, thus significantly reducing expert labor. Moreover, in domain specific instances such as EMRs, manually annotated logical forms allow the experts to express information essential for natural language understanding such as domain knowledge, temporal relations, and negation (Gao et al.; Chabierski et al., 2017). This knowledge, once captured, can be used to generate QA pairs on new documents, making the framework scalable.

4 Generating the emrQA Dataset

We apply the proposed framework to generate the emrQA corpus consisting of questions posed by physicians against longitudinal EMRs of a patient, using annotations provided by i2b2 (Figure 2).

4.1 Question Collection and Normalization

We collect questions for EMR QA by, 1) polling physicians at the Veterans Administration for what they frequently want to know from the EMR (976 questions), 2) using an existing source of 5,696 questions generated by a team of medical experts from 71 patient records (Raghavan et al., 2017) and 3) using 15 prototypical questions from an ob-

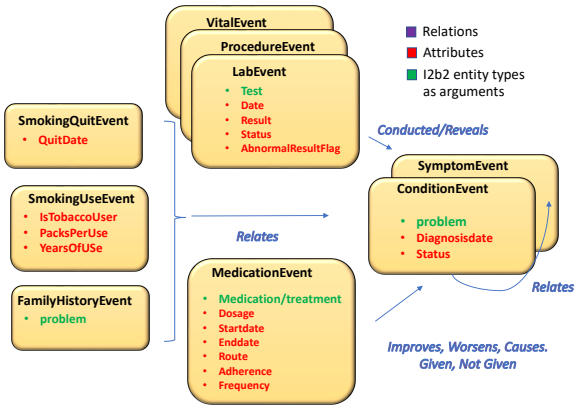

Figure 3: Events, attributes & relations in emrQA's logical forms. Events & attributes accept i2b2 entities as arguments.

servational study done by physicians (Tang et al., 1994). To obtain templates, the questions were automatically normalized by identifying medical entities (using MetaMap (Aronson, 2001)) in questions and replacing them with generic placeholders. The resulting ∼2K noisy templates were expert reviewed and corrected (to account for any entity recognition errors by MetaMap). We align our entity types to those defined in the i2b2 concept extraction tasks (Uzuner et al., 2010a, 2011) - *problem, test, treatment, mode and medication*. E.g., The question *What is the dosage of insulin?* from the collection gets converted to the template *What is the dosage of* $|medication|$*?* as shown in Fig.2. This process resulted in 680 question templates. We do not correct for the usage/spelling errors in these templates, such as usage of "pt" for "patient", or make the templates gender neutral in order to provide a true representation of physicians' questions. Further, analyzing these templates shows that physicians most frequently ask about test results (11%), medications for problem (9%), and problem existence (8%). The long tail following this includes questions about medication dosage, response to treatment, medication duration, prescription date, etiology, etc. Temporal constraints were frequently imposed on questions related to tests, problem diagnosis and medication start/stop.

4.2 Associating Templates w/ Logical Forms

The 680 question templates were annotated by a physician with their corresponding logical form (LF) templates, which resulted in 94 unique LF templates. More than one question template that map to the same LF are considered paraphrases of each other and correspond to a particular question type (Table 2). Logical forms are defined based on an ontology schema designed by medical experts (Figure 3). This schema captures entities in unstructured clinical notes through medical events and their attributes, interconnected through relations. We align the entity and relation types of i2b2 to this schema.

A formal representation of the LF grammar using this schema (Figure 3) is as follows. Medical events are denoted as ME_i (e.g LabEvent, ConditionEvent) and relations are denoted as RE_i (e.g conducted/reveals). Now, $ME[a_1, .., a_j, .., oper(a_n)]$ is a medical event where a_j represents the attribute of the event (such as result in LabEvent). An event may optionally include constraints on attributes captured by an operator ($oper() \in$ sort, range, check for null values, compare). These operators sometimes require values from external medical KB (indicated by *ref*, e.g. lab.*ref* low/lab.*ref* high to indicate range of reference standards considered healthy in lab results) indicating the need for medical knowledge to answer the question. Using these constructs, a LF can be defined using the following rules,

$LF \rightarrow ME_i \mid M_1$ relation M_2
$M_1 \rightarrow ME_i, M_2 \rightarrow ME_j$
$M_1 \rightarrow M_1$ relation $M_2, M_2 \rightarrow M_1$ relation M_2
$relation \rightarrow OR \mid AND \mid RE_i$

Advantages of our LF representation include the ability to represent composite relations, define attributes for medical events and constrain the attributes to precisely capture the information need in the question. While these can be achieved using different methods that combine lambda calculus and first order logic (Roberts and Demner-Fushman, 2016), our representation is more human comprehensible. This allows a physician to consider an ontology like Figure 3 and easily define a logical form. Some example question templates with their LF annotations are described in Table 3 using the above notation. The LF representation of the question in Figure 2 is *MedicationEvent(*$|medication|$*) [dosage=x]*. The entities seen in LF are the entities posed in the question and entity marked x indicates the answer entity type.

4.3 Template Filling and Answer Extraction

The next step in the process is to populate the question and logical form (QL) templates with existing annotations in the i2b2 clinical datasets and extract answer evidence for the questions.

Property	Example Annotation	Stats.
Fine grained answer type (attribute entity is answer)	Q: What is the **dosage** of $\|medication\|$? LF: MedicationEvent ($\|medication\|$) [**dosage=x**]	62.7%
Course grained answer type (event entity is answer)	Q: **What** does the patient take $\|medication\|$ for? LF: MedicationEvent($\|medication\|$)given{ConditionEvent(**x**) OR SymptomEvent(**x**)}	52.1%
Questions with operators on entities	Q: What are the last set of labs with **elevated** numbers out of range? LF: LabEvent (x) [date=x, (result=x)>lab.refhigh]	25.5%
Questions which require medical KB	Q: What are the last set of labs with **elevated numbers out of range**? LF: LabEvent (x) [date=x, (result=x)>**lab.refhigh**]	11.7%
At least one event relation	What lab results does he have that are pertinent to $\|problem\|$ **diagnosis** LF: LabEvent (x) [date=x, result=x] **conducted/reveals** ConditionEvent ($\|problem\|$)	46.8%

Table 3: Properties of question templates inferred from the corresponding logical form templates. The boldface words hint at the presence of the corresponding property in both question and the logical form template.

The i2b2 datasets are expert annotated with fine-grained annotations (Guo et al., 2006) that were developed for various shared NLP challenge tasks, including (1) smoking status classification (Uzuner et al., 2008), (2) diagnosis of obesity and its co-morbidities (Uzuner, 2009), extraction of (3) medication concepts (Uzuner et al., 2010a), (4) relations, concepts, assertions (Uzuner et al., 2010b, 2011) (5) co-reference resolution (Uzuner et al., 2012) and (6) heart disease risk factor identification (Stubbs and Uzuner, 2015). In Figure 2, this would correspond to leveraging annotations from medications challenge between medications and their dosages, such as *medication=Nitroglycerin, dosage=40mg*, to populate |medication| and generate several instances of the question "What is the dosage of $|medication|$?" and its corresponding logical form *MedicationEvent($|medication|$)[dosage=x]*. The answer would be derived from the value of the dosage entity in the dataset.

Preprocessing: The i2b2 entities are preprocessed before using them with our templates to ensure syntactic correctness of the generated questions. The pre-processing steps are designed based on the i2b2 annotations syntax guidelines (Guo et al., 2006). To estimate grammatical correctness, we randomly sampled 500 generated questions and found that <5% had errors. These errors include, among others, incorrect usage of article with the entity and incorrect entity phrasing.

Answer Extraction: The final step in the process is generating answer evidence corresponding to each question. The answers in emrQA are defined differently; instead of a single word or phrase we provide the entire i2b2 annotation line from the clinical note as the answer. This is because the context in which the answer entity or phrase is mentioned is extremely important in clinical decision making (Demner-Fushman et al., 2009). Hence, we call them answer evidence instead of just answers. For example, consider the question *Is the patient's $hypertension$ controlled?*. The answer to this question is not a simple *yes/no* since the status of the patient's $hypertension$ can change through the course of treatment. The answer evidence to this question in emrQA are multiple lines across the longitudinal notes that reflect this potentially changing status of the patients condition, e.g. *Hypertension-borderline today*. Additionally, for questions seeking specific answers we also provide the corresponding answer entities.

The overall process for answer evidence generation was vetted by a physician. Here is a brief overview of how the different i2b2 datasets were used in generating answers. The *relations challenge* datasets have various event-relation annotations across single/multiple lines in a clinical note. We used a combination of one or more of these, to generate answers for a question; in doing so we used the annotations provided by the *i2b2 co-reference* datasets. Similarly, the *medications challenge* dataset has various event-attribute annotations but since this dataset is not provided with co-reference annotations, it is currently not possible to combine all valid answers. The *heart disease challenge dataset* has longitudinal notes (∼5 per patient) with record dates. The events in this dataset are also provided with time annotations and are rich in quantitative entities. This dataset was primarily used to answer questions that require temporal and arithmetic reasoning on events. The patient records in the *smoking and obesity challenge* datasets are categorized into classes with no entity annotations. Thus, for questions generated on these datasets, the entire document acts as evidence and the annotated class information (7 classes) needs to be predicted as the answer.

The total questions, LFs and answers gener-

ated using this framework are summarized in Table 1. Consider the question *How much does the patient smoke?* for which we do not have i2b2-annotations to provide an answer. In cases where the answer entity is empty, we only generate the question and LF, resulting in more question types being used for QL than QA pairs: only 53% of question types have answers.

5 emrQA Dataset Analysis

We analyze the complexity of emrQA by considering the LFs for question characteristics, variations in paraphrases, and the type of reasoning required for answering questions (Table 2, 3, 4).

5.1 Question/Logical Form Characteristics

A quantitative and qualitative analysis of emrQA question templates is shown in Table 3, where logical forms help formalize their characteristics (Su et al., 2016). Questions may request specific fine-grained information (attribute values like dosage) or may express a more coarse-grained need (event entities like medications etc), or a combination of both. 25% of questions require complex operators (e.g compare(>)) and 12% of questions express the need for external medical knowledge (e.g. lab.refhigh). The questions in emrQA are highly compositional, where 47% of question templates have at least one event relation.

5.2 Paraphrase Complexity Analysis

Questions templates that map to the same LF are considered paraphrases (e.g, Table 2) and correspond to the same question type. In emrQA, an average of 7 paraphrase templates exist per question type. This is representative of FAQ types that are perhaps more important to the physician. Good paraphrases are lexically dissimilar to each other (Chen and Dolan, 2011). In order to understand the lexical variation within our paraphrases, we randomly select a question from the list of paraphrases as a reference and evaluate the others with respect to the reference, and report the average BLEU (0.74 ± 0.06) and Jaccard Score (0.72 ± 0.19). The low BLEU and Jaccard score with large standard deviation indicates the lexical diversity captured by emrQA's paraphrases (Papineni et al., 2002; Niwattanakul et al., 2013).

5.3 Answer Evidence Analysis

33% of the questions in emrQA have more than one answer evidence, with the number ranging from 2 to 61. E.g., the question *Medications Record?* has all medications in the patient's longitudinal record as answer evidence. In order to analyze the reasoning required to answer emrQA questions, we sampled 35 clinical notes from the corpus and analyzed 3 random questions per note by manually labeling them with the categories described in Table 4. Categories are not mutually exclusive: a single example can fall into multiple categories. We compare and contrast this analysis with SQuAD (Rajpurkar et al., 2016), a popular MC dataset generated through crowdsourcing, to show that the framework is capable of generating a corpus as representative and even more complex. Compared to SQuAD, emrQA offers two new reasoning categories, temporal and arithmetic which make up 31% of the dataset. Additionally, over two times as many questions in emrQA require reasoning over multiple sentences. Long and noisy documents make the question answering task more difficult (Joshi et al., 2017). EMRs are inherently noisy and hence 29% have incomplete context and the document length is 27 times more than SQuAD which offers new challenges to existing QA models. Owing to the domain specific nature of the task, 39% of the examples required some form of medical/world knowledge.

As discussed in Section 4.3, 12% of the questions in emrQA corpus require a class category from *i2b2 smoking and obesity datasets* to be predicted. We also found 6% of the questions had other possible answers that were not included by emrQA, this is because of the lack of co-reference annotations for the *medications challenge*.

6 Baseline Methods

We implement baseline models using neural and heuristic methods for question to logical form (Q-L) and question to answer (Q-A) mapping.

6.1 Q-L Mapping

Heuristic Models: We use a template-matching approach where we first split the data into train/test sets, and then normalize questions in the test set into templates by replacing entities with placeholders. The templates are then scored against the ground truth templates of the questions in the train set, to find the best match. The placeholders in the LF template corresponding to the best matched question template is then filled with the normalized entities to obtain the predicted LF. To normalize the test questions we use CLiNER

Reasoning	Description	Example Annotation	emrQA	SQuAD
Lexical Variation (Synonym)	Major correspondence between the question and answer sentence are synonyms.	Q: Has this patient ever been **treated** with insulin? E: Patient sugars were **managed** o/n with sliding scale insulin and diabetic	15.2%	33.3%
Lexical Variation (world/medical knowledge)	Major correspondence between the question and answer sentence requires world/medical knowledge to resolve	Q: Has the patient complained of any **CAD symptoms**? E: 70-year-old female who comes in with ***substernal chest pressure***	39.0%	9.1%
Syntactic Variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence	Q: Has this patient ever been **treated** with ffp? E: attempt to reverse anticoagulation , one unit of **FFP was begun**	60.0%	64.1%
Multiple Sentence	Co-reference and higher level fusion of multiple sentences	Q: What happened when the patient was given **ascending aortic root replacement**? E: The patient tolerated **the procedure** fairly well and was transferred to the ICU with his *chest open*	23.8%	13.6%
Arithmetic	Knowing comparison and subtraction operators.	Q: Show me any **LDL** $>$ **100 mg/dl** in the last 6 years? E: gluc 192, ***LDL 115***, TG 71, HDL 36	13.3%	N.A.
Temporal	Reasoning based on time frame	Q: What were the results of the abnormal A1C on **2115-12-14**? E: HBA1C **12/14/2115** *11.80*	18.1%	N.A.
Incomplete Context	Unstructured clinical text is noisy and may have missing context	Q: What is her current dose of iron? E: Iron *325 mg* p.o. t.i.d.	28.6%	N.A.
Class Prediction	Questions for which a specific predefined class needs to be predicted	Q: Is the patient currently Obese? E: Yes	12.4%	N.A.

Table 4: We manually labeled 105 examples into one or more of the above categories. Words relevant to the corresponding reasoning type are in bold and the answer entity (if any) in the evidence is in *italics*. We compare this analysis with SQuAD.

Dataset	Train/Test	HM-1	HM-2	Neural
GeoQuery	600/280	32.8%	52.1%	74.6%[†]
ATIS	4,473/448	20.8%	52.2%	69.9%[†]
emrQL-1	1M/253K	0.3%	26.3%	22.4%
emrQL-2	1.1M/296K	31.6%	32.0%	42.7%

Table 5: Heuristic (HM) and neural (seq2seq) models performance on question to logical form learning in emrQA.

(Boag et al., 2015) for emrQA and Jia and Liang (2016)'s work for ATIS and GeoQuery. Scoring and matching is done using two heuristics: (1) HM-1, which computes an identical match, and (2) HM-2, which generates a GloVe vector (Arora et al., 2016) representation of the templates using sentence2vec and then computes pairwise cosine similarity.

Neural Model: We train a sequence-to-sequence (seq2seq) (Sutskever et al., 2014) with attention paradigm (Bahdanau et al., 2014; Luong et al., 2017) as our neural baseline (2 layers, each with 64 hidden units). The same setting when used with Geoquery and ATIS gives poor results because the parameters are not appropriate for the nature of that dataset. Hence, for comparison with GeoQuery and ATIS, we use the results of seq2seq model with a single 200 hidden units layer (Jia and Liang, 2016). At test time we automatically balance missing right parentheses.

[†]results from Jia and Liang (2016)

6.1.1 Experimental Setup

We randomly partition the QL pairs in the dataset in train(80%) and test(20%) sets in two ways. (1) In emrQL-1, we first split the paraphrase templates corresponding to a single LF template into train and test, and then generate the instances of QL pairs. (2) In emrQL-2, we first generate the instances of QL pairs from the templates and then distribute them into train and test sets. As a result, emrQL-1 has more lexical variation between train and test distribution compared to emrQL-2, resulting in increased paraphrase complexity. We use accuracy i.e, the total number of logical forms predicted correctly as a metric to evaluate our model.

6.1.2 Results

The performance of the proposed models is summarized in Table 5. emrQL results are not directly comparable with GeoQuery and ATIS because of the differences in the lexicon and tools available for the domains. However, it helps us establish that QL learning in emrQA is non-trivial and supports significant future work.

Error analysis of heuristic models on emrQL-1 and emrQL-2 showed that 70% of the errors occurred because of incorrect question normalization. In fact, 30% of these questions had not been normalized at all. This shows that the entities

added to the templates are complex and diverse and make the inverse process of template generation non trivial. This makes a challenging QL corpus that cannot trivially be solved by template matching based approaches.

Errors made by the neural model on both emrQL-1 and emrQL-2 are due to long LFs (20%) and incorrectly identified entities (10%), which are harder for the attention-based model (Jia and Liang, 2016). The increased paraphrase complexity in emrQL-1 compared to emrQL-2 resulted in 20% more structural errors in emrQL-1, where the predicted event/grammar structure deviates significantly from the ground truth. This shows that the model is not adequately capturing the semantics in the questions to generalize to new paraphrases. Therefore, emrQL-1 can be used to benchmark QL models robust to paraphrasing.

6.2 Q-A Mapping

Question-answering on emrQA consists of two different tasks, (1) extraction of answer line from the clinical note (machine comprehension (MC)) and (2) prediction of answer class based on the entire clinical note. We provide baseline models to illustrate the complexity in doing both these tasks.

Machine Comprehension: To do extractive QA on EMRs, we use DrQA's (Chen et al., 2017) document reader which is a multi-layer RNN based MC model. We use their best performing settings trained for SQuAD data using Glove vectors (300 dim-840B).

Class Prediction: We build a multi-class logistic regression model for predicting a class as an answer based on the patient's clinical note. Features input to the classifier are TF-IDF vectors of the question and the clinical notes taken from *i2b2 smoking and obesity datasets*.

6.2.1 Experimental setup

We consider a 80-20 split of the data for train-test. In order to evaluate worst-case performance, we train on question-evidence pairs in a clinical note obtained by using only one random paraphrase for a question instead of all the paraphrases. We use a slightly modified[‡] version of the two popularly reported metrics in MC for evaluation since our evidence span is longer: Exact Match (EM) and F1. Wherever the answer entity in an evidence is explicitly known, EM checks if the answer entity is present within the evidence, otherwise it checks if the predicted evidence span lies within ± 20 characters of the ground truth evidence. For F1 we construct a bag of tokens for each evidence string and measure the F1 score of the overlap between the two bags of tokens. Since there may be multiple evidence for a given question, we consider only the top 10 predictions and report an average of EM and F1 over ground truth number of answers. In the class prediction setting, we report the subset accuracy.

Model	Train/Test	Exact Match	F1
DrQA (MC)	47,605/9,966	59.2%	60.6
Class Prediction	1276/320	36.6%	n.a

Table 7: Performance of baseline models on the two QA sub tasks, machine comprehension (MC) and class prediction.

6.2.2 Results

The performance of the proposed models is summarized in Table 7. DrQA is one of the best performing models on SQuAD with an F1 of 78.8 and EM of 69.5. The relatively low performance of the models on emrQA (60.6 F1 and 59.2 EM) shows that QA on EMRs is a complex task and offers new challenges to existing QA models.

To understand model performance, we macro-average the EM across all the questions corresponding to a LF template. We observe that LFs representing temporal and arithmetic[§] needs had $< 16\%$ EM. LFs expressing the need for medical KB[§] performed poorly since we used general Glove embeddings. An analysis of LFs which had approximately equal number of QA pair representation in the test set revealed an interesting relation between the model performance and LF complexity, as summarized in Table 6. The trend shows that performance is worse on multiple relation questions as compared to single relation and attribute questions, showing that the LFs sufficiently capture the complexity of the questions and give us an ability to do a qualitative model analysis.

Error analysis on a random sample of 50 questions containing at least one answer entity in an evidence showed that: (1) 38% of the examples required multiple sentence reasoning of which 16% were due to a missing evidence in a multiple evidence question, (2) 14% were due to syntactic variation, (3) 10% required medical reasoning and (4) in 14%, DrQA predicted an incomplete evidence span missing the answer entity in it.

[‡] using the original definitions, the evaluated values were far less than those obtained in Table 7

[§] maximum representation of these templates comes from the i2b2 heart disease risk dataset

Logical Form template	Property	Exact Match
MedicationEvent ($\|medication\|$) [**enddate=x**]	single attribute	55.3%
{LabEvent ($\|test\|$) OR ProcedureEvent ($\|test\|$)} **conducted** {ConditionEvent(x) OR SymptomEvent (x)}	single relation	32.2%
{MedicationEvent($\|treatment\|$)ORProcedureEvent($\|treatment\|$)} **improves/worsens/causes** {ConditionEvent (x) OR SymptomEvent (x)}	multiple relation	12.6%

Table 6: Neural models (DrQA) performance on question-evidence corpus of emrQA stratified according to the logical form templates. Instance showing increasing complexity in the logical forms with decreasing model performance.

7 Discussion

In this section, we describe how our generation framework may also be applied to generate open-domain QA datasets given the availability of other NLP resources. We also discuss possible extensions of the framework to increase the complexity of the generated datasets.

Open domain QA dataset generation: Consider the popularly used SQuAD (Rajpurkar et al., 2016) reading comprehension dataset generated by crowdworkers, where the answer to every question is a segment of text from the corresponding passage in the Wikipedia article. This dataset can easily be generated or extended using our proposed framework with existing NLP annotations on Wikipedia (Auer et al., 2007; Nothman et al., 2008; Ghaddar and Langlais, 2017).

For instance, consider DBPedia (Auer et al., 2007), an existing dataset of entities and their relations extracted from Wikipedia. It also has its own ontology which can serve as the semantic frames schema to define logical forms. Using these resources, our reverse engineering technique for QA dataset generation can be applied as follows. (1) Question templates can be defined for each entity type and relation in DBPedia. For example[¶], consider the relation *[place, country]* field in DBpedia. For this we can define a question template *In what country is |place| located?*. (2) Every such question template can be annotated with a logical form template using existing DBPedia ontology. (3) By considering the entity values of DBPedia fields such as *[place=Normandy, dbo:country=France]*, we can automatically generate the question *In what country is Normandy located?* and its corresponding logical form from the templates. The text span of *country=France* from the Wikipedia passage is then used as the answer (Daiber et al., 2013). Currently, this QA pair instance is a part of the SQuAD dev set. Using our framework we can generate many more instances like this example from different Wikipedia passages - without crowdsourcing efforts.

[¶]example reference: http://dbpedia.org/page/Normandy

Extensions to the framework: The complexity of the generated dataset can be further extended as follows. (1) We can use a coreferred or a lexical variant of the original entity in the question-logical form generation. This can allow for increased lexical variation between the question and answer line entities in the passage. (2) It is possible to combine two or more question templates to make compositional questions with the answers to these questions similarly combined. This can also result in more multiple sentence reasoning questions. (3) We can generate questions with entities not related to the context in the passage. This can increase empty answer questions in the dataset, resulting in increased negative training examples.

8 Conclusions and Future Work

We propose a novel framework that can generate a large-scale QA dataset using existing resources and minimal expert input. This has the potential to make a huge impact in domains like medicine, where obtaining manual QA annotations is tedious and infeasible. We apply this framework to generate a large scale EMR QA corpus (emrQA), consisting of 400,000 question-answers pairs and 1 million question-logical forms, and analyze the complexity of the dataset to show its non-trivial nature. We show that the logical forms provide a symbolic representation that is very useful for corpus generation and for model analysis. The logical forms also provide an opportunity to build interpretable systems by perhaps jointly (or latently) learning the logical form and answer for a question. In future, this framework may be applied to also re-purpose and integrate other NLP datasets such as MIMIC and generate a more diverse and representative EMR QA corpus (Johnson et al., 2016).

Acknowledgments

This project is partially funded by Sloan Research Fellowship, PhRMA Foundation Award in Informatics, and NSF Career Award (1652815). The authors would like to thank Siddharth Patwardhan for his valuable feedback in formatting the paper.

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In Proceedings of the AMIA Symposium, page 17. American Medical Informatics Association.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In The semantic web, pages 722–735. Springer.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. 2014. Modeling biological processes for reading comprehension. In EMNLP.

William Boag, Kevin Wacome, Tristan Naumann, and Anna Rumshisky. 2015. Cliner: A lightweight tool for clinical named entity recognition. AMIA Joint Summits on Clinical Research Informatics (poster).

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl_1):D267–D270.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. arXiv preprint arXiv:1506.02075.

Qingqing Cai and Alexander Yates. 2013. Large-scale semantic parsing via schema matching and lexicon extension. In ACL (1), pages 423–433.

Piotr Chabierski, Alessandra Russo, and Mark Law. 2017. Logic-based approach to machine comprehension of text.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051.

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 190–200. Association for Computational Linguistics.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In Proceedings of the 9th International Conference on Semantic Systems (I-Semantics).

Randall Davis, Bruce Buchanan, and Edward Shortliffe. 1977. Production rules as a representation for a knowledge-based consultation program. Artificial intelligence, 8(1):15–45.

Dina Demner-Fushman, Wendy Webber Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics, 42(5):760–772.

Mary Devereaux. 2013. The use of patient records (ehr) for research.

Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. 2017. Quasar: Datasets for question answering by search and reading. arXiv preprint arXiv:1707.03904.

Minwei Feng, Bing Xiang, Michael R Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on, pages 813–820. IEEE.

Yansong Feng, Songfang Huang, Dongyan Zhao, et al. 2016. Hybrid question answering over knowledge base and free text. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2397–2407.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. 2010. Building watson: An overview of the deepqa project. AI magazine, 31(3):59–79.

Jianfeng Gao, Rangan Majumder, and Bill Dolan. Machine reading for question answering: from symbolic to neural computation.

Abbas Ghaddar and Phillippe Langlais. 2017. Winer: A wikipedia annotated corpus for named entity recognition. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 413–422.

Y. Guo, R. Gaizauskas, I. Roberts, G. Demetriou, and M. Hepple. 2006. Identifying personal health information using support vector machines. In i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, pages 10–11.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems, pages 1693–1701.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. arXiv preprint arXiv:1606.03622.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. Scientific data, 3:160035.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551.

Harlan M Krumholz, Sharon F Terry, and Joanne Waldstreicher. 2016. Data acquisition, curation, and use for a continuously learning health system. Jama, 316(16):1669–1670.

Chonho Lee, Zhaojing Luo, Kee Yuan Ngiam, Meihui Zhang, Kaiping Zheng, Gang Chen, Beng Chin Ooi, and Wei Luen James Yip. 2017. Big healthcare data analytics: Challenges and applications. In Handbook of Large-Scale Distributed Computing in Smart Healthcare, pages 11–41. Springer.

Vanessa Lopez, Christina Unger, Philipp Cimiano, and Enrico Motta. 2013. Evaluating question answering over linked data. Web Semantics: Science, Services and Agents on the World Wide Web, 21:3–13.

Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. 2017. Neural machine translation (seq2seq) tutorial. https://github.com/tensorflow/nmt.

George A Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In Proceedings of the International MultiConference of Engineers and Computer Scientists, volume 1.

Joel Nothman, James R Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. In Proceedings of the Australasian Language Technology Association Workshop 2008, pages 124–132.

Hamid Palangi, Paul Smolensky, Xiaodong He, and Li Deng. 2018. Question-answering with grammatically-interpretable representations. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

Preethi Raghavan and Siddharth Patwardhan. 2016. Question answering on electronic medical records. In Proceedings of the 2016 Summit on Clinical Research Informatics, San Francisco, CA, March 2016.

Preethi Raghavan, Siddharth Patwardhan, Jennifer J. Liang, and Murthy V. Devarakonda. 2017. Annotating electronic medical records for question answering. arXiv:1805.06816.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In EMNLP, volume 3, page 4.

Kirk Roberts and Dina Demner-Fushman. 2016. Annotating logical forms for ehr questions. In LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation, volume 2016, page 3772. NIH Public Access.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. IEEE Journal of Biomedical and Health Informatics.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. Journal of biomedical informatics, 58:S20–S29.

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gur, Zenghui Yan, and Xifeng Yan. 2016. On generating characteristic-rich question sets for qa evaluation. In EMNLP, pages 562–572.

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. arXiv preprint arXiv:1703.10090.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112.

Paul C Tang, Danielle Fafchamps, and Edward H Shortliffe. 1994. Traditional medical records as a source of clinical data in the outpatient setting. In Proceedings of the Annual Symposium on Computer Application in Medical Care, page 575. American Medical Informatics Association.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. arXiv preprint arXiv:1611.09830.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia

Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics, 16(1):138.

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, and Christina Unger. 2015. Hawk–hybrid question answering using linked data. In European Semantic Web Conference, pages 353–368. Springer.

Özlem Uzuner. 2009. Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association, 16(4):561–570.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. Journal of the American Medical Informatics Association, 19(5):786–791.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association, 15(1):14–24.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010a. Extracting medication information from clinical text. Journal of the American Medical Informatics Association, 17(5):514–518.

Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010b. Community annotation experiment for ground truth generation for the i2b2 medication challenge. Journal of the American Medical Informatics Association, 17(5):519–523.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18(5):552–556.

Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. 2012. Making machine learning models interpretable. In ESANN, volume 12, pages 163–172. Citeseer.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In Trec, volume 99, pages 77–82.