EHR Question Answering and Timeline Generation: Methods for Large-scale Corpus and Model Creation

MIT PIs: Peter Szolovits
IBM PIs: Preethi Raghavan
Research Project Term: Start Date: 08/01/2018, Estimated Completion: 08/01/2021
Pillars: AI Algorithms; Application of AI to Industries
Type of proposal: Core

Keywords: Natural language processing; question answering; temporal relation learning; unstructured clinical notes; corpus creation; semi- supervised learning; event progressions; medical AI

Synopsis: The is a dire need for large-scale annotated corpora and standardized credible benchmarks exploring state-of-the-art deep learning methods in the clinical domain. The lack of these resources prevents training of supervised models and evaluating approaches, thus significantly stunting clinical NLP research. The challenges involving the tedious and fine-grained nature of clinical annotation tasks, need for domain-expertise, and personal health identifiers in the EHR, prevent both easy large-scale annotation, as well as sharing of EHR data. In light of these challenges, we focus on minimizing annotator burden, creating large-scale datasets, and training state-of-the-art models in the context of two critical NLP problems: (1) Question Answering and (2) Temporal Relation Learning. For (1), 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 community-shared QA dataset for electronic medical records by leveraging existing expert annotations on clinical notes for various NLP tasks from existing datasets like i2b2 [1], MIMIC[2]. We characterize the dataset and explore its learning potential by training neural models for question to logical form and answer mapping. In case of (2), we propose the idea of capturing the chronological order medical events as they naturally occur across patients and creating an event progression KB. Specifically, we explore re-purposing large-scale structured data available from various sources to create event progressions. Here, we propose a novel methodology to first generate partial timelines from a patient's (unstructured and structured) EHR and automatically completing the timeline using the event progression KB. We will design, implement and evaluate the timelines models in the context of several medical AI applications.

1 Introduction

Physicians frequently seek answers to questions from unstructured electronic health records (EHRs) to support clinical decision-making (Demner-Fushman et al., 2009; Simpson et al., 2014; Tang et al., 1994). But, in a significant majority of cases, they are unable to find the information they want from EHRs (Tang et al., 1994). Natural language processing (NLP) has the ability to change this, by transforming, improving and potentially revolutionizing the way physicians access information about a patient and make decisions about their treatment. Electronic health records (EHRs) document the healthcare provided to the patient using both unstructured clinical notes and structured data. To navigate the vast amounts of information about a patient, current EHR applications available to physicians either attempt to proactively summarize important patient information or they provide a more traditional search service that retrieves notes that match query terms. However, it may be possible to much more precisely (and automatically) answer very specific questions about the patient's health, by providing short answers to natural language questions – perhaps supported by evidence. For e.g., "Were there any complications of the patient's RYGB surgery?" could be answered with specific related phrases or passages from the EHR (Figure 1).

[1] https://www.i2b2.org/NLP/DataSets/
[2] https://mimic.physionet.org/

While there have been several attempts at question answering in the NLP community ((Rajpurkar et al., 2016; Voorhees et al., 1999; Ferrucci et al., 2013)), there have been no question answering efforts in the clinical domain. One reason for this is that patient-specific QA from an EHR has significantly different challenges when compared to open-domain QA. While EHR QA may be likened to machine comprehension ((Rajpurkar et al., 2016; Gao et al., 2017)), it is complicated by challenges brought on by little to no redundancy in facts, and longitudinal, temporal and domain-specific nature of information centered around a patient. Moreover, the nature of questions is not always factoid. Therefore, deeper analysis of clinical text is required to address problems like temporal reasoning, relation detection, discourse analysis both within and across clinical notes (Raghavan and Patwardhan, 2016). Another important reason is that there are no community-shared datasets (of any scale) for question answering on EHR data.

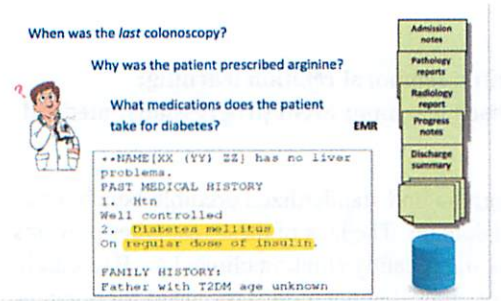

Figure 1: Physicians seek answers to questions from unstructured clinical notes and structured data for decision making.

The lack of large-scale community-shared annotations pervades all problems in clinical NLP, making it difficult to train credible, representative models and take advantage of advances in artificial neural networks. Consider the problem of temporal relation learning to generate a medical event timeline from a patient's EHR (Figure **??**). Medical events are temporally-associated concepts in clinical text that include medical conditions affecting the patient's health, or procedures, tests performed on a patient. Having access to a timeline of such events is fundamental to understanding clinical narratives and key to applications such as longitudinal studies, question answering and document summarization (Zhou and Hripcsak, 2007). However, annotating fine-grained relations between events, event at a document-level, is very tedious. Thus, even the popularly used general-domain newswire corpus, Timebank (Pustejovsky et al., 2003), is rather small-scale in the number of annotated temporal relations. In the medical domain, this is further complicated by multiple data sources (structured and unstructured) and longitudinal notes that mention events that go back and forth in time. **Importantly, the lack of large-scale annotated datasets prevents the research community from benchmarking methods, results and making collective progress towards solving clinical NLP problems.** This is especially problematic because it is complex and difficult to use existing open-domain NLP tools on EHR data (Demner-Fushman et al., 2009).

The reasons this are many fold: 1) Serious privacy concerns about personal health information (PHI) ((Guo et al., 2006)) rule out popular crowd-sourcing options. 2) Annotating clinical notes requires domain expertise; often, these annotations must be done by physicians themselves. This is time-consuming, tedious and impractical, given the scale and detail required in the annotations for most tasks. Thus, very few datasets like i2b2 challenge datasets ((Uzuner et al., 2010b)), MIMIC ((Johnson et al., 2016)) (developed over several years in collaboration with large medical groups and hospitals) share small-scale annotated clinical notes for certain clinical NLP tasks. This proposal addresses both 1) and 2) in the context of two complex and essential NLP problems in clinical text - question answering and automatic medical event timeline generation from clinical notes.

The overall theme of this proposal tries to answer the following questions: (1) How can we re-purpose resources that already exist, minimize manual annotation efforts, and generate credible large-community-shared datasets to train usable models in the clinical domain? (2) How can we ease the expert annotation effort in generating gold-standard annotations for fine-grained and challenging clinical NLP problems? Our goals are as follows.

- Create the **first-ever community-shared large-scale dataset for question answering on EHRs** that includes questions, logical forms (symbolic representation of the question), and answer evidence in the context of a clinical note.
- Explore the complexity and the learning potential of the corpus by **training baseline as well as improved neural models on the corpus and make it available to the research community**.
- Develop novel methods to **minimize the cognitive burden on the expert while annotating clinical notes for fundamental NLP tasks, such as temporal relation learning, by creating a knowledge-**

base of medical event progressions that capture the natural order in which events may occur for different problems.

EHR Question Answering. The first part focuses on generating the first-ever large-scale dataset for physician question answering against the EHR (**emrQA**) that is representative of prototypical questions posed by physicians and captures the complexities that go into answering them in terms of logical forms and answer evidence. 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 EHR) (Lee et al., 2017). Thus, we propose a QA dataset generation process (Figure 3) that involves capturing physician information needs expressed as natural language questions from various sources, normalizing the questions to templates by replacing entities in the question with placeholders and annotating logical forms for the question templates. This is followed by populating the entity placeholders in the questions, logical forms and generating answer evidence using existing expert-annotated datasets (i2b2, MIMIC) for various other NLP tasks.

Logical forms are defined as an intermediate representation between the questions and the expected answer. They are symbolic representations that use relations from an ontology/schema to represent the relations in the question, and also associate the question information need with an answer entity types. Prior work has created logical forms using predicate calculus, first-order logic in the context of semantic parsing((Berant et al., 2013; Yih et al., 2016)). However, the ability to create logical forms that are easy to understand by the domain expert is valuable in closed domains where the answering needs are complex ((Roberts and Demner-Fushman, 2016)). Consider the question, "Does the patient take any medication for $|problem|$?". Here, $|problem|$ is a placeholder for any disease or symptom the patient suffers from and the answer must capture any current medication the patient is on for the $|problem|$. Now to accurately answer this, the system must find medications that treat $|problem|$ and where the end date of the medication is either after the current date (or does not exist). We want to capture this precise information in the logical form, thus connecting the physician's information need to the appropriate answer in the EHR. This may be captured in a logical form as follows:

"**MedicationEvent** (x) CheckIfNull ([enddate]) *treats* **ConditionEvent (|problem|)** *OR* **SymptomEvent (|problem|)** *OR* **MedicationEvent** (x) [enddate>currentDate] *treats* **ConditionEvent (|problem|)** *OR* **SymptomEvent (|problem|)**"

Here, the **events** (and its attributes and operators on them), *relations* between events are all grounded in a simple ontology developed by a domain-expert for clinical notes in the EHR. A snapshot of this is shown in Figure 4. Generating both logical forms and answers for a question allows users to build explainable EHR QA models learn both jointly learning them.

Moreover, 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, manually annotated logical forms allow experts to express information essential for natural language understanding such as domain knowledge, temporal relations and negation (Gao et al., 2017; Chabierski et al., 2017). This knowledge, once captured, can be used to generate QA pairs on new documents, making the framework scalable. We hypothesize that the proposed framework can generate a QA dataset for any new domain where the answering needs are complex. However, this framework, when applied to generate a corpus for EHR QA, would be particularly promising for medical AI by providing access to information required for active decision support locked in clinical notes in a patient's EHR.

Automated Timeline Completion Using Event Progressions. The next part of the proposal focuses on easing the annotation process for the task of temporal relation learning (Mani et al., 2006) between medical concepts in unstructured clinical notes. In doing this, we also propose a novel method for timeline generation from across structured and unstructured sources of a patient's record. We plan to achieve this by learning problem-specific event progressions, capturing the natural order of events. We will primarily generate these progressions from large-scale structured data across patient's in the EHR. An event progression is a sequence of medical events (problems, symptoms, tests, medications, procedures etc.) that have been observed frequently across patients that suffer from a specific problem. The specific problem maybe determined by the discharge (dx) code associated with the structured data (in case of outpatient records, the

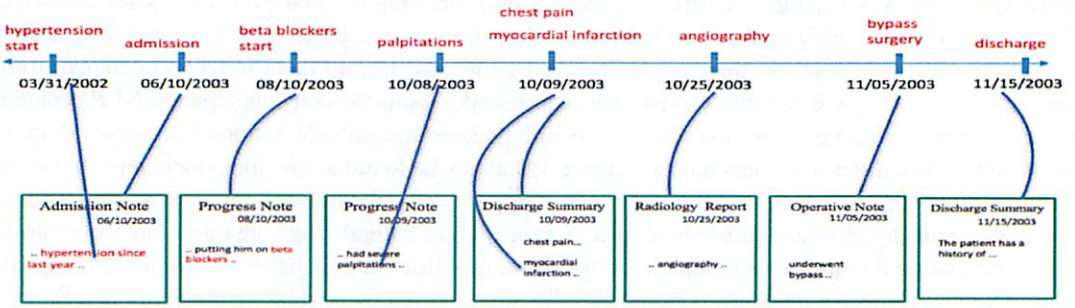

Figure 2: Timeline of chronologically ordered medical events found in unstructured clinical notes of a patient.

dx codes are accurate and is linked to each encounter). A problem and related medications, tests, procedures may also be derived from the schema of the structured data, if any. However, even in cases, where it is not possible to derive problem-specific progression, simply extracting frequent event sequences from across patient's provides valuable information about how medical events may be ordered for patient's with a certain mix of problems. We will also explore expanding event progressions with additional information by relating the problems in the structured data with additional concepts (such as symptoms) found in guidelines/ care plans used in evidence-based medicine or in general medical literature (journals, Wikipedia). We plan to utilize the event progressions to automatically complete partial timelines generated from across unstructured clinical notes in the EHR.

Often, there is limited information available in the text to assign timestamps to events and place them on the timeline; e.g., admission and discharge dates in clinical notes. Thus, only the subset of events that can be associated with explicit dates in the text are placed on the timeline leading to an incomplete partial timeline. The "unassigned events" are typically the multiple medical events within and across clinical notes that don't have explicit date/time anchors. The challenge here is given such a partial timeline generated from unstructured data, and events unassigned to the timeline, automatically place these unassigned events in appropriate positions on the timeline using an event progression KB. This is achieved via a mapping function that finds concepts in each event progression that are synonymous to the ones in the partial timeline and using them as anchor points to find the best progressions that are likely to match our current patient's timeline. We will then try to find the best alignment between the unassigned set and events from the selected event progressions via a match and align algorithm described in Section 2. We will also explore modeling the problem without trying to nail each event to a specific time point/ window by modeling the event sequences as a partial order graph or introducing a form of constraint reasoning using lower and upper bounds on time intervals between pairs of events (Kohane, 1987).

Additionally, we will also explore the possibility of adding semantic information to the event progression KB. This can again be achieved by utilizing any known semantic relations (treats, evaluates, causes, improves, worsens etc.) (Uzuner et al., 2011) between medical events expressed in the large-scale structured dataset (and also in other sources like medical literature). These relations will be rich in contextual relations observed across patients and serve as an informative prior in semantically relating events in clinical notes.

2 Research Plan

Our ultimate goal is to design general methodologies to create large-scale datasets for all clinical NLP problems with minimal expert input. In cases where expert input cannot be completely eliminated, we want to reduce the annotator burden by using semi-supervised learning methods like active learning. Importantly, we want the researchers in the community to make shared progress towards solving clinical NLP problems by providing shared large-scale datasets and baseline models as a foundation.

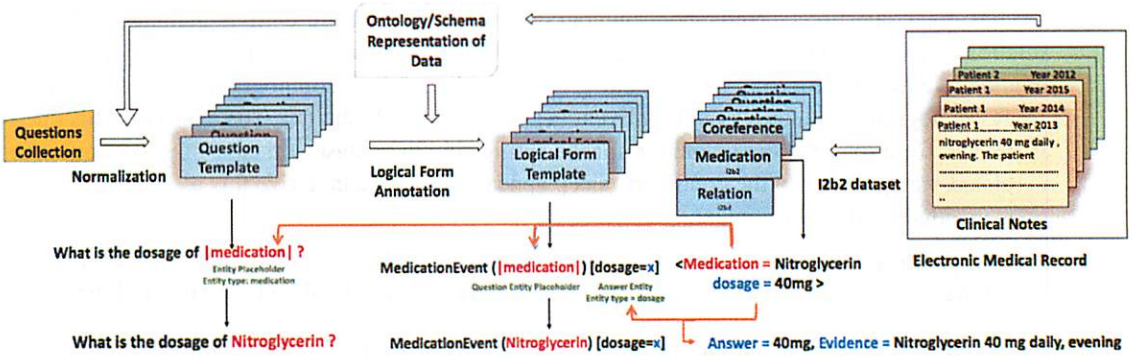

Figure 3: QA dataset generation framework using existing i2b2/ MIMIC annotations on a given patients record, to generate question, logical form and answer evidence. The highlights in the figure show the annotations being used for this example.

2.1 EMR Question Answering

- **Dataset generation:** Here, we create a large-scale corpus (emrQA) of questions, corresponding logical forms, and answers using minimal expert input. In doing this, we re-purpose existing annotations in community-shared datasets i2b2 (Uzuner et al., 2010b) and MIMIC (Johnson et al., 2016). The proposed generation process is shown in Figure 3.

  Question Collection: We collect questions by polling physicians for frequently asked questions against the EMR, both retrospectively and at the time of care. We currently have ≈7000 questions from the following sources. 1) Physicians at the Veterans Administration (VA) were polled for what they frequently want to know from the EMR (976 questions), 2) Questions generated by a team of medical experts on a set of patient records from Cleveland Clinic (5,696 questions) (Raghavan, 2017) and 3) Prototypical questions from an observational study done by physicians (Tang et al., 1994) (15 questions). An example of a frequently question from the VA: "Why did the patient have a colonoscopy?". We will additionally poll physicians within IBM (from different specialties) to further diversify the question set. The objective here is to capture the natural distribution of such a dataset by collecting questions from different expert-sources.

- **Question Templates:** To obtain templates, the questions will be automatically normalized by identifying medical entities (using MetaMap (Aronson, 2001), CliNER (Boag et al., 2015), cTakes (Savova et al., 2010)) and replacing them with generic placeholders. The resulting noisy templates will be expert reviewed and corrected (to account for any entity recognition errors by entity recognizer). The example VA question would be normalized to "Why did the patient have a $|test|$?" by replacing the entity colonoscopy with its semantic type placeholder $|test|$.

- **Question Logical Forms:** The question templates will then be annotated by a physician with their corresponding logical form templates. We develop a logical form representation using a ontology (a snapshot is seen in Figure 4)) that captures events and relations in unstructured clinical notes. We will align the entity and relation types of i2b2 and MIMIC to this schema. The corresponding logical form for the e.g. template is LabEvent (|testl) OR ProcedureEvent (|testl) evaluates ConditionEvent (x) OR SymptomEvent (x) that captures the information need expressed in the questions and links it to what is expected in the answer. In this case, the logical form tries to find the condition or symptom that colonoscopy was used to evaluate. We define a logical form grammar grounded in an ontology defined specifically to represent events and relations in unstructured EHR data (e.g., Figure 4). We also ensure that the logical forms are more human comprehensible so that it is easier for a physician to annotate logical forms for question templates.

- **Populating question and logical form templates with existing annotations in the i2b2/MIMIC datasets and extracting answer evidence.** The i2b2 datasets are expert annotated with fine-grained annotations (Guo et al., 2006) that were developed for various shared NLP challenge tasks including extracting entities, medications, coreference, relations, and heart disease risk factors. This may gives us "Why did the patient have a polysomnogram?", "Why did the patient have a brain MRI?" among

several other questions. Here, polysomnogram and brain MRI are entities annotated in the i2b2 relations challenge dataset. Further, we populate the logical form by again replacing ltestl with the annotated entity from i2b2. This would give us LabEvent (polysomnogram) OR ProcedureEvent (polysomnogram) evaluates ConditionEvent (x) OR SymptomEvent (x). The answer to this question is derived from the relation annotation for the entities in i2b2 dataset. In case of polysomnogram, it is related via evaluates to the condition "sleep apnea". This we can derive this as the answer and the text surrounding it as the context.

In incorporating a dataset like MIMIC, we could use the logical forms to query the structured data for answers and have a more complete emrQA dataset over structured and unstructured records.

- **EHR QA Models.** We will train a recurrent neural network (seq2seq models) with attention and copy architecture (Liang, 2016) for learning to map questions to logical forms. The copy is particularly important as it helps in ensuring that words from the question can be copied to populate placeholders in the logical forms. For the question answering model, we plan to build a system that is similar in architecture to a machine comprehension system where we are searching for an answer to a question in a large corpus of unstructured clinical notes for a patient. In order to achieve this, we will train a machine comprehension model as a multi-layer recurrent neural network that tries to find the answer to any question as text spans in one or more of the returned clinical notes. Further, the logical forms also provide an opportunity to build interpretable systems by perhaps jointly learning the logical form and answer for a given question. This will provide a rationale behind the extracted answer for a question. The rationale is credible as the logical forms for question templates were annotated by physicians themselves.

Our exploratory study with the i2b2 dataset shows us that we can generate approximately 1 million questions-logical form and 400,000+ question-answer evidence pairs.

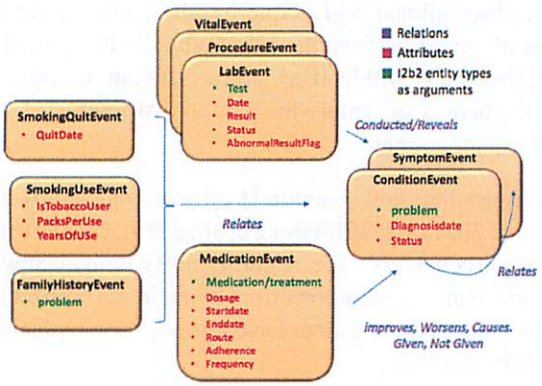

Figure 4: Events, attributes & relations in emrQA logical forms. Events & attributes accept i2b2 entities as arguments.

Additionally, we also explore some of the other opportunities that the dataset presents. While the question collection is already rich in paraphrases, a possible way of adding diversity to the question templates is by taking the templates and crowd-sourcing (putting them on Turk) them for more linguistic variants and incomplete "Google query" ways of asking questions. We also explore generating implicit sub-questions for each question. For example, when the physician asks: Consider the frequently-asked physician question, "What happened as the result of a treatment?" Here, the physician wants to know the outcome of a prescribed course of treatment, i.e., a procedure, medication or a lifestyle change. However, for efficient decision making, she would also need to know if the treatment improved/ worsened the condition, led to new problems, side effects, whether the treatment was adhered to and tolerated. These sub-questions maybe be created by linking relevant question templates. They may also be further expanded with some expert input. The sub-question generation may also be complementary to learning the right amount of context required to answer different types of questions. E.g., certain questions may need a note as the answer or several notes, or passages across notes, or a couple of sentences in a note or simply a phrase. This is determined both by the information need expressed in the question and the type of answer that may be useful to clinical decision making.

Evaluation. The emrQA will be the first-ever community shared large scale EMR QA dataset. We will evaluate the dataset using baseline models to determine its learning potential and complexity. We will also demonstrate through several baselines why QA systems developed on open source data (like Wikipedia) do not perform well on EMR data. We will compare the architectural differences in the neural models used for EMR QA vis-a-vis other popular open-domain QA systems.

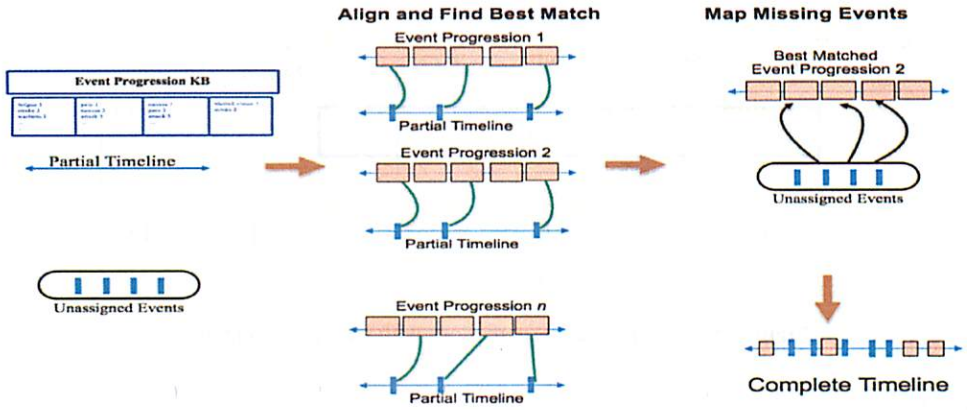

Figure 5: Proposed Match and Align methodology for automated timeline completion

2.2 Learning Event Progressions to Help Timeline Generation and Relation Learning

The main problems addressed here is learning problem-specific event progressions, creating an event progression KB, and generating medical event timelines across unstructured and structured data for a patient in the EHR. An event progression captures the natural sequence of events for a patient with a problem (say myocardial infarction/ set of problems (myocardial infarction, hypertension, diabetes). It may be possible to constrain events in the progression to ones that are related to the myocardial infarction (if those relations are explicit in the structured data). In the more general case, we wouldn't constrain the events, with the assumption that the structured data being large-scale would be representative of the mix of problems and related conditions that patients who suffers from mycoardial infarction typically also suffer from. In order to facilitate the mapping of unassigned events into the timeline using event progressions, for each event, we capture the following time information: Rank information implicit in the ordering, approximate distance from next event, confidence or fuzziness factor for the distance. Thus, it is possible to accommodate all temporal relations (before, after, overlaps, simultaneous, begins with, end with), with a degree of uncertainty, in mapping unassigned events to the partial timeline. This uncertainty helps capture missing information that may not allow us to ground an event in a precise instance of time accurately. Once we have the the event progression KB, the challenge is to find the event progression that best matches and aligns with a set of unassigned events and place those events in the partial timeline (generated based on events with explicit timestamps for that patient).

This can be achieved using the process described below (Figure 6).

Match and Align: A mapping function finds the event progressions in the KB that best match the partial timeline (Figure 6). The function outputs a similarity score between every 2 events across the event progression and partial timeline. E.g., a simple mapping function would output a score of 1 for identical and synonymous events and 0 otherwise. $Map(\text{myocardial infarction}_{event-progression}, \text{heart attack}_{partial-timeline}) = 1$ $Map(\text{dizziness}_{event-progression}, \text{chest pain}_{partial-timeline}) = 0$ The similarity of events is calculated using a knowledge-based approach by leveraging ontologies like UMLS or Wordnet. We then leverage the scores produced by the mapping function to align all event progressions against the partial timeline. For alignment, we will first use popular dynamic programming algorithms like Needleman Wunsch or Smith Waterman and pick the highest scoring alignment as the best match. Alternately, we will also develop a framework where we could pick the top n alignments as our best matched event progressions. **Timeline Completion:** Here, we match events from the missing event set to the selected event progression. This is achieved by mapping the matched missing events to the partial timeline by leveraging temporal information in the event progression to establish the position and date of the missing events in the partial timeline. We will develop a mapping function to achieve this based on the semantic similarity or synonymity between event pairs. For each unassigned event mapped to the event progression, we leverage available time information to estimate the relative position of the missing event in the partial timeline. We also check rank and distance of events in the progression and estimate a position and score for placing each missing event on the partial timeline based on its relative rank and distance to the next event using the

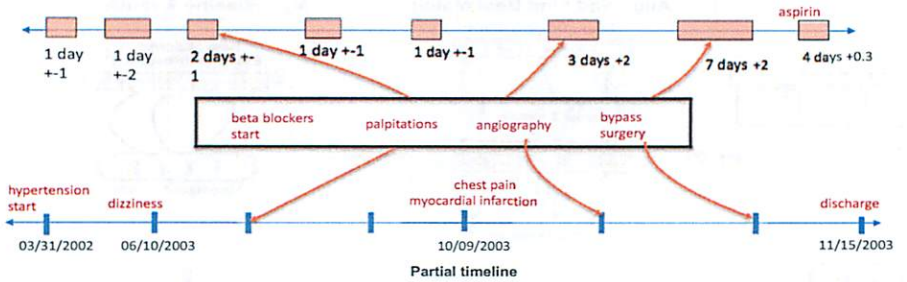

Figure 6: Matching event progressions with the partial timeline

following algorithm.

PositionFinder(Relative Rank, Distance, Confidence Interval) = Score

1. Say the unassigned event *"palpitation"* maps to *"palpitation"* in the selected event progression
2. From the event progression, we learn that palpitation occurs 2 days after dizziness, and 1 day before chest pain
3. Using this time information, we find, say, dizziness and chest pain on the partial timeline and place palpitations approximately 2 days after dizziness and 1 day before chest pain
4. Repeat this process for all unassigned events

We will also explore machine learning methods that incorporate features obtained by knowledge-driven and distributional methods to calculate similarity. This will help extend the matching and alignment problem as follows: All permutations of missing events are placed in the partial timeline to generate multiple candidate timelines. We can then align and score each candidate with each event progression and pick the alignment pair of (Event Progression, Candidate) with maximum score. We could also pick the top n aligned event progressions and use information from across these n progressions to complete the partial timeline.

Evaluation. We will be generating two usable resources—the event progressions KB and the completed timelines on datasets like MIMIC. The EMR datasets used for automated partial timeline completion will have some events that are timestamped (perhaps in structured data) and unassigned events (from the unstructured data). The event progressions are representative of frequent sequences of events present in the structured data. As long as physicians have entered correct concept names and timestamps in the structured tables, the progressions are bound to be accurately representative of the population they are created from. Since manually evaluating the quality of the timelines is as challenging as generating gold-standard annotations for the task, we propose the following evaluation methods.

1) Physicians will manually evaluate a sampled subset of the event progressions. They will be presented with a minimal set of temporally related pairs from each note (and some across notes) and required to place the events on the generated timeline.

2) Indirect evaluation by using the information from the timeline to inform in several other clinical NLP tasks that require this temporal information. Some examples of such tasks are learning about what was "planned" as the result of an assessment, "what happened" as the result of a plan. These tasks require the model to know that a certain event may be a consequence of another event; hence temporal ordering matters. We will measure the performance of these tasks (currently part of the Gemstone project) after including information extracted from these timelines.

3) Indirect evaluation by using it for the EMR QA task. We will evaluate the performance of question answering on the generated emrQA corpus, by using the timelines to help answer questions, especially those with temporal constraints.

3 Statement of Work

- **August 2018 - December 2018:** emrQA version 1 development and release using i2b2 data. Baseline, state-of-the art neural models for QL and QA learning.
- **October 2018 - January 2019:** Expanding emrQA to i2b2+MIMIC. Crowd sourcing the paraphrase generation task to introduce more variance in the way questions are expressed.
- **December 2018 - July 2019:** Design and develop models for additional tasks supported by the emrQA dataset. Question similarity or paraphrasing, sub-question generation.
- **October 2018 - February 2019:** Research on training the best explainable AI model using the emrQA dataset. We will jointly optimizing both the logical form and answer for a particular question in a multi-layered RNN framework.
- **November 2018 - March 2019:** Generate an event progression KB using structured MIMIC, Cleveland Clinic and Explorys.
- **January 2018 - April 2019:** Enrich the event progression KB with related concepts and semantic information from external sources.
- **March 2018 - July 2019:** Create timelines from MIMIC EHR data by leveraging the event progression KB.
- **December 2019 onwards:** Work on evaluations, both direct and indirect, to enable sharing updated versions of emrQA and MIMIC timelines.

Publication venues: ACL, EMNLP, NAACL, NIPS, TACL, JAMIA, JBI.

MIT PIs: Peter Szolovits (15%)
MIT RAs: RA1 (100%), UG RA2
IBM PIs: Preethi Raghavan(30%)
IBM Advisors: Ching-Huei Tsou, Uri Kartoun

4 Ambitious Objective

Scientific goals: Develop explainable QA models for physician question answering against the EMR. This increases credibility and usability of the model in a practical clinical setting.
Technological goals: Deploy a system trained on emrQA in a clinical setting such as a hospital or clinic, perhaps alongside / or integrated with the EHR software that is being used such as Epic.
Societal goals: Allow physicians access to nuanced information buried in unstructured EHRs and ease clinical decision making; thus overall improving the quality of healthcare.
Long-range impact: Provide the clinical NLP community with standard large-scale annotated datasets sufficiently representative of domain-specific challenges, such that models trained on it would work across institutions. Only with access to such corpora can the community progress and make shared efforts towards truly understanding the unstructured EHR automatically.

5 Confidential Information

We have an IBM patent related to the second problem in the proposal i.e. Automated Timeline Completion Using an Event Progression Knowledge Base (US Patent App. 15/172,813, 2017).

6 Data, including any unique Data conditions

1. The community-shared 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 it 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).

2. NLP annotations created on MIMIC data as part of various SemEval and CLEF/ShARe challenges (Roberts et al., 2007).

3. Structured data available in MIMIC and IBM-internal structured data available from Cleveland Clinic and Explorys.

4. Other publicly shared NLP corpora to generate a QA dataset for other domains using our QA dataset generation process.

7 Software

1. emrQA dataset - a large scale corpus for question answering against the EMR. Community-shared, subject to the same license agreement that the datasets used to generate emrQA (i2b2 and MIMIC for now) are under. The scripts to generate emrQA will be made open source (anyone with i2b2 or MIMIC license should be able to generate emrQA).

2. Baseline models benchmarking performance on emrQA for various tasks like question-logical form learning, question-answering, jointly learning questions-logical forms-answers and question paraphrase learning will be made open source. This will allow researchers to use, explore and improve on our models and provide feedback to improve the quality of the dataset.

3. Medical event timelines generated from across unstructured and structured notes in MIMIC will be made available subject to the same license agreement as MIMIC.

4. The event progressions estimated from MIMIC data alone will be released under the same license as MIMIC - we will release scripts to generate these from MIMIC data. The complete event progression KB that also uses IBM-internal structured data cannot be shared under the current agreement with the providers. This can be revisited later as making the event progressions available would immensely benefit several clinical NLP tasks.

8 Budget

PI Name: Peter Szolovits
Sponsor: IBM
Title: TBD
Period: 09/01/18 - 08/31/21

	# OF MONTH	09/01/18 08/31/19	09/01/19 08/31/20	09/01/20 08/31/21	GRAND TOTAL
PERSONNEL					
Peter Szolovits	0.67 MOS	12,064	10,178	10,649	32,891
RA PhD (1)	12 MOS	42,375	43,647	44,956	130,978
UROP (1)	12 MOS	10,370	10,681	11,001	32,052
Total Salaries & Wages		64,809	64,506	66,607	195,921
Technical and Admin Support - Not MTDC Base		6,183	6,156	6,076	18,416
Employee Benefits		3,088	2,606	2,726	8,420
Employee Benefits - Not MTDC Base		1,583	1,576	1,556	4,715
Vacation Accrual - Not MTDC Base		470	468	462	1,400
Sub-Total of Fringe Benefits		5,141	4,650	4,744	14,534
TOTAL PERSONNEL COSTS		76,133	75,312	77,427	228,872
OPERATING EXPENSES					
M & S		3,000	3,000	0	6,000
Service Centers		1,668	1,668	1,668	5,004
RA Tuition - Not MTDC Base		25,760	27,048	28,400	81,208
Allocated Expenses (Year 1 - 0.90%) - Not MTDC Base		625	623	614	1,862
Sub-Total of Tuition and M&S Allocation		26,385	27,671	29,015	83,071
TOTAL OPERATING EXPENSES		31,053	32,339	30,683	94,075
TOTAL DIRECT COSTS		107,187	107,650	108,110	322,946
OVERHEAD (F&A)		42,813	42,350	41,890	127,054
TOTAL PROPOSAL COSTS		150,000	150,000	150,000	450,000
MTDC Base		72,565	71,779	71,001	215,345
Allocation Base		69,477	69,174	68,275	206,925

References

Aronson, A. R. (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.

Berant, J., Chou, A., Frostig, R., and Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In EMNLP, volume 2, page 6.

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

Chabierski, P., Russo, A., and Law, M. (2017). Logic-based approach to machine comprehension of text.

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

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., and Mueller, E. T. (2013). Watson: beyond jeopardy! Artificial Intelligence, 199:93–105.

Gao, J., Majumder, R., and Dolan, B. (2017). Machine reading for question answering: from symbolic to neural computation.

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

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. Scientific data, 3:160035.

Kohane, I. S. (1987). Temporal reasoning in medical expert systems. Technical report, Boston Univ., MA (USA).

Lee, C., Luo, Z., Ngiam, K. Y., Zhang, M., Zheng, K., Chen, G., Ooi, B. C., and Yip, W. L. J. (2017). Big healthcare data analytics: Challenges and applications. In Handbook of Large-Scale Distributed Computing in Smart Healthcare, pages 11–41. Springer.

Liang, P. (2016). Learning executable semantic parsers for natural language understanding. Communications of the ACM, 59(9):68–76.

Mani, I., Verhagen, M., Wellner, B., Lee, C. M., and Pustejovsky, J. (2006). Machine learning of temporal relations. In ACL.

Pustejovsky, J., Castaño, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). TimeML: Robust specification of event and temporal expressions in text. In New Directions in Question Answering'03, pages 28–34.

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

Raghavan, Preethi; Patwardhan, S. L. J. J. D. M. V. (2017). Annotating electronic medical records for question answering. arXiv:1805.06816.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A., et al. (2007). The clef corpus: semantic annotation of clinical text. In AMIA Annual Symposium Proceedings, volume 2007, page 625. American Medical Informatics Association.

Roberts, K. and Demner-Fushman, D. (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.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Schuler, K. K., and Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. JAMIA, pages 507–513.

Simpson, M. S., Voorhees, E. M., and Hersh, W. (2014). Overview of the trec 2014 clinical decision support track. Technical report, LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD.

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

Tang, P. C., Fafchamps, D., and Shortliffe, E. H. (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.

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

Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., and South, B. R. (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.

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

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

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

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

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

Yih, W.-t., Richardson, M., Meek, C., Chang, M.-W., and Suh, J. (2016). The value of semantic parse labeling for knowledge base question answering. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 201–206.

Zhou, L. and Hripcsak, G. (2007). Temporal reasoning with medical data—a review with emphasis on medical natural language processing. Journal of biomedical informatics, 40(2):183–202.

9 Biosketch

9.1 Peter Szolovits

Peter Szolovits is Professor of Computer Science and Engineering in the MIT Department of Electrical Engineering and Computer Science (EECS) and an Associate faculty member in the MIT Institute of Medical Engineering and Science (IMES) and its Harvard/MIT Health Sciences and Technology (HST) program. He is also head of the Clinical Decision-Making Group within the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). His research centers on the application of AI methods to problems of medical decision making, natural language processing to extract meaningful data from clinical narratives, and the design of information systems for health care institutions and patients. He has worked on problems of diagnosis, therapy planning, execution and monitoring for various medical conditions, computational aspects of genetic counseling, controlled sharing of health information, privacy and confidentiality issues in medical record systems, and integration of clinical and genomic data for translational medicine. His interests in AI include knowledge representation, qualitative reasoning, probabilistic inference, and machine learning. He has supervised 35 doctoral theses, been a member of over 50 more doctoral thesis committees, supervised 73 Master's theses in these research areas, and served as a mentor to over a dozen postdoctoral and medical Fellows. Much of the work of his lab focuses on the use of natural language processing methods to extract facts from clinical narratives and on the analysis of clinical and genomic data. A complete list of his publications can be found here.

9.2 Preethi Raghavan

Preethi Raghavan is a Research Staff Member at IBM working on natural language processing problems in clinical text. Her current research focuses on developing question answering technologies to help physicians interact with a patient's EHR and extract the information they need to make well-informed decisions about a patient's healthcare. Over the past 10 years, she has worked on several natural language processing and medical AI problems such as temporal relation learning, coreference resolution, textual entailment and published in several top NLP conferences like ACL, EMNLP, and NAACL. She obtained a PhD in Computer Science from The Ohio State University in 2014 (won a best dissertation proposal award at AMIA). A complete list of her publications can be found here.