Annotating Logical Forms for EHR Questions

Kirk Roberts[†], **Dina Demner-Fushman**[‡]
† School of Biomedical Informatics
University of Texas Health Science Center at Houston
Houston TX, USA
kirk.roberts@uth.tmc.edu

‡ Lister Hill National Center for Biomedical Communications
National Library of Medicine
National Institutes of Health
Bethesda MD, USA
ddemner@mail.nih.gov

Abstract

This paper discusses the creation of a semantically annotated corpus of questions about patient data in electronic health records (EHRs). The goal is to provide the training data necessary for semantic parsers to automatically convert EHR questions into a structured query. A layered annotation strategy is used which mirrors a typical natural language processing (NLP) pipeline. First, questions are syntactically analyzed to identify multi-part questions. Second, medical concepts are recognized and normalized to a clinical ontology. Finally, logical forms are created using a lambda calculus representation. We use a corpus of 446 questions asking for patient-specific information. From these, 468 specific questions are found containing 259 unique medical concepts and requiring 53 unique predicates to represent the logical forms. We further present detailed characteristics of the corpus, including inter-annotator agreement results, and describe the challenges automatic NLP systems will face on this task.

Keywords: question answering, semantic parsing, electronic health records

1. Introduction

Over the past few years, the adoption of electronic health records (EHRs) has grown remarkably in the United States (Office of the National Coordinator for Health Information Technology, 2014) and many other developed countries as well (Neumann, 2010). However, many usability issues with EHRs are a barrier to their effective use (Zhang and Walji, 2014), including difficulty accessing the information stored in EHRs. Natural language question answering (QA) provides an intuitive interface for retrieving EHR data by reducing the need to understand the internal organization of the data. However, since this data is stored in both unstructured text and structured databases, a deep semantic understanding of EHR questions is necessary for an effective QA system.

Consider the following questions:

(1) What is her blood pressure and how low has it been?

(2) Is he on any inotropes?

(3) Is she wheezing this morning?

(4) Who is the patient's nephrologist?

For a complete semantic understanding of these questions, several types of linguistic processing must be performed. First, Question (1) must be decomposed into two distinct questions. Second, medical concepts ("*blood pressure*", "*inotropes*", "*wheezing*", and "*nephrologist*") must be recognized and normalized to an ontology consistent with how the data is stored in EHRs. Third, the phrase "*this morning*" must be recognized as an explicit temporal constraint on the answer space. Finally, the questions must be converted to a logical form that captures their full meaning. This includes many other linguistic tasks, such as the implicit temporal semantics in each question. For the questions above, that meaning could be represented by the following lambda calculus forms:

(1a') $\textit{latest}(\lambda x.\textit{has\_test}(x, \texttt{0005823}, \texttt{visit}))$

(1b') $\textit{min}(\lambda x.\textit{has\_test}(x, \texttt{0005823}, \texttt{visit}))$

(2') $\delta(\lambda x.\textit{has\_treatment}(x, \texttt{0304509}, \texttt{status}))$

(3') $\delta(\lambda x.\textit{has\_problem}(x, \texttt{0043144}, \texttt{status}) \wedge \textit{time\_within}(x, \text{`this morning'}))$

(4') $\textit{latest}(\lambda x.\textit{has\_doctor}(x, \texttt{0260039}, \texttt{history}))$

These logical forms will be explained in more detail in Sections 3-5. After semantic parsing, a QA system can map the logical forms to structured queries (such as SQL for relational databases) or particular NLP-based extractors.

In order to develop supervised semantic parsers capable of performing this deep semantic analysis automatically, a training corpus of question/logical form pairs is necessary. This paper describes the process of creating such a corpus from 446 EHR questions collected by Li (2012). To simplify the semantic parsing task, we provide two additional layers of annotations. First, a syntactic decomposition layer (Section 3) breaks multi-part questions, such as Question (1), to ensure each sub-question has a single answer. Second, a normalization layer (Section 4) maps clinical concepts to an ontology and performs various types of shallow semantic processing. The key idea is to simplify

the question as much as possible prior to input to the semantic parser. Finally, a logical form layer indicates the final semantic form (Section 5).

The corpus described in this paper was manually double-annotated with each of these layers and reconciled. Section 6 describes the results of the annotation, including corpus analysis and inter-annotator agreement. Section 7 discusses some of the linguistic challenges with this task.

2. Background

Medical Question Answering: Medical QA has seen significant interest (Athenikos and Han, 2010) due to the tremendous amount of biomedical knowledge, far beyond what any one clinician or researcher could comprehend. The field of medical QA has largely focused on searching for information outside the EHR, both targeted toward clinicians (Yu and Sable, 2005; Kobayashi and Shyu, 2006; Demner-Fushman and Lin, 2007; Schardt et al., 2007; Terol et al., 2007; Yu and Cao, 2008; Athenikos et al., 2009; Cairns et al., 2011; Cao et al., 2011; Patrick and Li, 2012) and consumers (Zhang, 2010; Liu et al., 2011; Andersen et al., 2012; Kilicoglu et al., 2013; Van Der Volgen et al., 2013; Roberts et al., 2014b; Roberts et al., 2014c; Roberts et al., 2014a; Roberts and Demner-Fushman, 2016). Clinician-targeted QA systems typically focus on the biomedical literature, while consumer QA systems focus on consumer-friendly websites such as MedlinePlus[1] (Schnall and Fowler, 2013). Additional work in Information Retrieval (IR) has sought to bring relevant literature to clinicians using only a small set of general questions (Simpson et al., 2014; Roberts et al., 2015; Roberts et al., 2016).

Significantly less work has focused on QA for the EHR, though a good amount of attention has been paid to IR for the EHR (Voorhees and Tong, 2011; Voorhees and Hersh, 2012; Hanauer et al., 2014). Of note, Patrick and Li (2012) produced the set of EHR questions used here (Li, 2012). In their work, machine learning (ML) based text classification is used to identify a fixed number of templates, then information extraction ML recognizers are used to identify the arguments to these templates. Conversely, the logical forms presented here allow for a greater variety of questions to be recognized without pre-defining templates and customizing ML extractors for each template argument. The most similar work to our own, which understands questions about patient data using a logical form approach is Waldinger et al. (2011). While they focus on aggregate patient data instead of a specific patient (e.g., "*Find patients on a regimen containing EFV and 3TC.*"), this distinction is fairly minor. Instead, the main difference is the scope of their work is far more limited (only questions that can be answered by the Stanford HIV Drug Resistance Database), which enables semantic parsing to be performed using a small number of hand-built rules. To build a more generalizable QA system with logical forms, sufficient data is necessary to train state-of-the-art semantic parsers.

Semantic Parsing: Semantic parsers have received tremendous interest as of late. Much of the work can be organized by what type of data is used to train the parser. This includes logical forms (Zettlemoyer and Collins, 2005; Wong and Mooney, 2007; Muresan, 2011), question-answer pairs (Clarke et al., 2010; Liang et al., 2011), conversation logs (Artzi and Zettlemoyer, 2011), and even unsupervised semantic parsing (Poon, 2013). While most of these focused on semantic parsing of questions, recent work includes semantic parsing to Abstract Meaning Representation (Artzi et al., 2015). In this work, our goal is to provide a sufficient number of question/logical form pairs to train a baseline semantic parser. However, the broad range of the medical domain likely means that additional types of data will be necessary to achieve human-like semantic parsing capabilities for EHR questions.

3. Question Decomposition

The first annotation layer simplifies questions by splitting those that contain multiple sub-questions, a process we refer to as question decomposition (Roberts et al., 2014b). This ensures that every question has a single logical form that provides one specific answer (though this answer could be a set of answers). In the EHR question data, decomposition is rarely necessary but it is required on a handful of questions. Consider the following questions with their decompositions:

(5) Is she awake and obeying commands?
  (a) Is she awake?
  (b) Is she obeying commands?

(6) What is the pacing mode and underlying rhythm?
  (a) What is the pacing mode?
  (b) What is the underlying rhythm?

(7) What is the blood glucose and how high / low has it been?
  (a) What is the blood glucose?
  (b) How high has the blood glucose been?
  (c) How low has the blood glucose been?

We follow the procedures described in Roberts et al. (2014b) for syntactic decomposition and assembly of the decomposed questions. Most of the decomposable questions require breaking a coordination into two or more parts. Further, where relevant, such as Question (7), coreferential anaphors are resolved in the decomposed questions so that each sub-questions can be considered self-contained.

4. Normalization

The purpose of the normalization layer is to simplify the input to the semantic parser by leveraging existing NLP methods, including the recognition of medical concepts and temporal expressions. Consider the following questions and their normalizations:

(8) What is his emotional status?
→ What is `patient_pos nn:concept(C0684322)` ?

(9) What is her ventilation?
→ What is `patient_pos nn:concept(C2945579)` ?

[1] http://www.nlm.nih.gov/medlineplus/

(10) Did the patient's temperature exceed 38C in the last 48 hrs?
→ Did `patient_pos nn:concept(C0005903)` exceed `measurement`('38C') in `temporal_ref`('the last 48 hrs') ?

Several forms of normalization/recognition have been applied here to make the semantic parsing task easier. First, "*his*", "*her*", and "*the patient's*" have all been normalized to the single term `patient_pos`, meaning the parser won't need to learn equivalent partient references entirely from the data. Second, medical concepts have been normalized to a structured ontology and represented with an ID. In order for a semantic parser to do this straight from question/logical form pairs, many examples of every concept would need to be in the data (for reference, some medical ontologies contain over 100,000 concepts). Instead, automatic NLP methods exist to perform this task (Pradhan et al., 2015). Finally, other expressions like the temperature measurement "*38C*" and temporal expression "*the last 48 hrs*" are annotated. In this work, we do not normalize these expressions (this would require a grounding date for the temporal expressions, which we do not have). Instead, by recognizing them at the normalization layer, the semantic parser can simply pass their arguments to the equivalent logical form. For instance, `temporal_ref`('the last 48 hrs') would be mapped to the logical expression *time_within*(• , 'the last 48 hrs'), where the • argument is derived elsewhere.

We consider six different normalization categories for EHR questions:

- **Major Hospital Event**: `admission`, `discharge`, `readmission`
- **Major Hospital Relative Event**: `preoperative`, `intraoperative`, `postoperative`
- **Person Reference**: `patient`, `patient_pos`, `staff`
- **Measurement**: `measurement`(E), where E is some string referring to a quantitative measurement such as temperature, weight, or length
- **Spatial Reference**: `location_ref`(E), where E is some location string, typically referring to a place within the hospital (e.g., "*the ICU*")
- **Temporal Reference**: `temporal_ref`(E), where E is a string with an absolute or relative date/time
- **Medical Concept**: A reference to some clinically applicable concept, normalized to SNOMED-CT (Stearns et al., 2001), but using UMLS (Lindberg et al., 1993) identifiers to be consistent with other normalization tasks (Pradhan et al., 2015). The annotations take the form *POS*:*TYPE*(*ID*), where *POS* is the part-of-speech, and *TYPE* is a semantic type for the concept with the given *ID*. In this way, what we actually annotate looks slightly different than Questions (8)-(10), but is functionally equivalent. For example, when the concept corresponds to the UMLS type *disease* or *injury*, we use the label `problem`, and when the concept is a UMLS *finding*, we use the label `finding`. So in Question (8), we would actually use `nn:finding(C0684322)` instead of `nn:concept`. This is simply for human readability and annotation ease, and is similarly handled in the logical form annotation. When input to a semantic parser, all the labels would be collapsed into `concept` with the appropriate part-of-speech.

Predicate	Description
$\delta(S)$	Whether the set S is non-empty
at_location(a, τ)	Whether event a occurred at the location denoted by the string τ
greater_than(a, b), *less_than*(a, b)	Whether the value of event a is greater or less than the value of b
is_result(a, n)	Whether n is the result of event a
latest(S)	Returns the most recent event in set S
max(S), *min*(S)	Largest/smallest value in set S
positive(S), *negative*(S)	Filters the events in set S, returning only the positive/negative events (diagnostic tests, disease diagnoses, etc.)
sum(S)	Sum of the results in set S
time(a)	The time of event a
time_within(a, τ)	Whether event a is within the time denoted by the string τ

Table 1: Common non-concept predicates used in the EHR question corpus.

5. Logical Form

To represent the deep semantics of EHR questions, λ-calculus expressions are used. These combine first order logic expressions with λ-expressions that denote sets matching a particular condition. This corresponds well to the organization of structured queries and is similar to many other semantic parsing tasks (Zettlemoyer and Collins, 2005; Artzi et al., 2014). The logical forms in this corpus combine the quantifier λ, predicates (boolean predicates and functions), and variables and literals that act as arguments to the predicates. The predicates can be broken down into two main types: concept predicates that retrieve information from the EHR, and non-concept predicates that manipulate that information. The most common non-concept predicates and their descriptions are detailed in Table 1.

Concept predicates are boolean functions that take the form *has_TYPE*(*EVENT*, *CONCEPT*, *TIME*), where *TYPE* is the semantic type of the *CONCEPT* (the normalized ID), *EVENT* is a variable that refers to the event that is an instantiation of the concept, and *TIME* is the implicit timeframe for the event (explicit temporal constraints use the predicate *time_within*). Similar to normalization, the *TYPE* is just for readability, and at runtime all the concept predicates (which we refer to collectively as *has_**) are collapsed into a single predicate. Questions (1)-(4) and logical forms (1a')-(4') show examples of how the complete logical forms are made with these components.

The *TIME* argument is particularly interesting, as well as challenging. Since the *has_** predicates are used to retrieve events, we need to place a temporal restriction on

when those events could have occurred. Naively, one might assume all events of a certain type should be considered, but in general physicians are focused only on recent events. If they ask for the patient's blood work, but no test had been performed in the 5 years, it is probably better to return nothing since such dated results are of little use. Conversely, if we assume all unspecified events must have occurred within the hospital or doctor's visit, we would miss relevant events from the patient's history. Instead, the implied time is a combination of the semantics of the event as well as subtle linguistic clues in the question. We consider five possible times: `pmh` (past medical history), `history` (all history up to present), `visit` (current inpatient/outpatient visit), `status` (current status), and `plan` (future event). The overlap of these times are formally defined as:

$$\texttt{history} \supseteq \texttt{visit} \supseteq \texttt{status}$$
$$\texttt{history} \supseteq \texttt{pmh}$$
$$\texttt{pmh} \cap \texttt{visit} = \emptyset$$
$$\texttt{history} \cap \texttt{plan} = \emptyset$$

Question (1) indicates events that occurred during the current `visit`, while Questions (2) and (3) are concerned with the patient's `status`. The difference relates to the semantics of tests, treatments, and problems. Tests are generally short events, thus occurring in the near past, while problems are more interesting when they are active, and treatments can be either. This also corresponds to the way many EHRs store information. For example, problem lists store only active problems. However, Question (4) refers to the patient's doctor, which could easily be outside a hospital, and thus uses `history`.

6. Results

The 446 questions were double-annotated and reconciled by two experts in medical informatics. The annotation occurred layer-wise: annotate normalizations, reconcile normalizations, annotate logical forms, and then reconcile logical forms (syntactic decomposition is rarely needed and does not require multiple annotation). The first 100 questions were used for developing the logical forms and proceeded in three phases (first 25, then another 25, then the remaining 50). Finally, the full set was double-annotated layer-wise and reconciled.

From the 446 original questions, syntactic decomposition resulted in 468 sub-questions. From these, 420 concepts were normalized (259 unique). The most frequent concepts are shown in Table 2. Constructing the logical forms required 1,470 predicates (53 unique), the most common of which are shown in Table 3. Only four of the time arguments were annotated (`pmh` questions are certainly possible, but not within this data set), and were dominated by the `visit` time (72%), with most of the remainder being `status` (25%). The `plan` time was only 2% of the cases, while the `history` was only 0.5%

When measuring inter-annotator agreement, it is important to note that logical form annotations can be broken down into sub-annotations, so measuring agreement on complete logical forms is overly strict. The annotators were in complete agreement for 58% of questions at the normalization layer and 42% of questions at the logical form layer.

UMLS CUI	#	Concept Name
C0392201	12	Blood glucose measurement
C0456388	8	Blood product
C0005903	5	Body Temperature
C0021925	4	Intubation
C0013227	4	Pharmaceutical Preparations
C0445623	4	Microorganism
C1704353	4	fluid - substance
C0042036	4	Urine
C2076600	4	Influenza due to Influenza A virus subtype H1N1
C1320225	3	Adverse drug event resulting from treatment of disorder

Table 2: Ten most frequent concepts, including their UMLS Concept Unique Identifier (CUI), frequency, and preferred name.

Predicate	#
λ	443
*has_**	428
δ	216
latest	127
time_within	51
time	32
sum	25
min	14
max	12
at_location	12

Table 3: Ten most frequent predicates and their frequencies.

Agreement in the individual components was much greater. We measure the following using an F_1-measure where one set of annotations acts as the gold and the other acts as the guess annotations. The most difficult annotation was concept normalization, which had an F_1 of 0.61. This is due to the many possible concepts in SNOMED-CT that a word or phrase could be normalized to. Often, there were up to a dozen possible choices. During reconciliation, a semantic type heuristic was devised (see Discussion) which would have significantly improved agreement by providing a method for choosing between equally valid normalizations. Temporal expressions had better agreement (F_1 = 0.88), with the most disagreements relating to whether to consider simple temporal words (e.g., "*date*", "*time*"), which are no longer annotated. Similar disagreements occured for spatial references (0.75), which were rare. There was complete agreement on measurements, which were also rare.

Logical form annotation has two main components where disagreement is possible: predicates and implicit time arguments. All other arguments are inherited from the normalization layer. Predicate agreement was quite good (F_1 = 0.85). Agreement on specific predicates varied greatly. The agreement on λ-expressions was near perfect (0.98). Other common predicates were quite high, such as δ (0.95), the *has_** predicates (0.96), and *time_within* (0.95). However, agreement on the *latest* predicate was less impressive (0.64). The rarer predicates also tended to bring agreement down, with a combined F_1 of 0.68. More difficult than predicates was assigning the implicit time argu-

ment to the *has_** predicates. This had an F_1 agreement of 0.72. The difficulty here is in combining medical knowledge with assumptions of the author's intent. It was previously decided that, for the most part, the medical events are limited to within the `visit` timeframe, but most disagreements were between `visit` and `status`. As such, further assumptions were made based on the semantic type. For example, unless explicitly stated, tests and findings are assumed to have occurred in the recent past, and are thus `visit` instead of `status`.

7. Discussion

Concept normalization is known to be a difficult task due to UMLS ambiguity and many disambiguation methods have been proposed for automated systems(Jimeno-Yepes et al., 2011). Manual annotation revealed yet another layer of complexity: choosing the best semantic nuance. For example, in "*What is the blood glucose and how high / low has it been?*", the concept *blood glucose* could be normalized to a `procedure` (UMLS preferred name: "Glucose measurement, blood"), a `finding` ("Finding of blood glucose level"), or an `observable_entity` ("Blood glucose status, or Organic Chemical"). All these mappings are acceptable in the context, but we still needed to agree on one, which led to establishing a sequence for prioritizing mappings during the early reconciliation phase. For instance, when available a `substance` was prioritized over a `procedure`, and a `procedure` was prioritized over a `finding`. The prioritization rules are based on the observed richness of semantic relations in SNOMED CT that will allow most flexibility in reaching other related concepts.

Another difficulty in normalization was to decide how closely the mappings should follow the form of the question as opposed to the intended meaning. Although we tried our best to avoid inference and stick to the literal meaning when possible, some of the literal mappings would have been wrong. For instance, in "*What microorganisms were cultured?*" the concepts would have been `C0007635` ("Cultured cells") and `C0445623` ("Microorganism"), which will present difficulty generating query to find positive microbiology test results in the end-goal application. Therefore, C2242979 ("Microbial culture") appears to be a much better mapping for this question.

We aimed to create the smallest possible set of predicates to have more examples for each. For the most part, the questions can be captured by a handful of predicates. For several rare questions we chose representation with the existing predicates, perhaps losing some nuanced information, rather than creating a one-off new exact predicate. In those cases, we attempted to create a predicate that was simple to recognize, such as the *is_significant* predicate for the question "*Has there been a bleeding event and is it significant?*". The far greater difficulty in this question is determining whether an event is significant given the vast number of possible medical events. This, however, is not a natural language problem and is out of the scope of this work.

Finally, as can be seen by relatively low agreement on several sub-tasks, some parts of this task were non-trivial even for manual annotation, which indicates the tasks will be challenging for the current NLP systems, e.g., little work has been done so far on ranking adequate normalization options or deciding when to pick a pre-coordinated concept and when to chose post-coordination. The positive aspect of our work, however, is that when possible we tried to align difficult tasks in this work with existing clinical natural lanuguage processing tasks for which a separate large dataset is available, e.g., Pradhan et al. (2015).

8. Conclusion

We have described the creation of a corpus of EHR questions annotated with logical forms. Our goal is to provide semantic parsers with sufficient data to convert a natural language EHR query into a structured form that may be used to query an EHR database for patient information. The questions in the corpus were annotated at multiple layers: (a) syntactic decomposition, (b) concept normalization, and (c) lambda-calculus-based logical forms. Agreement results indicate that certain aspects of this task are quite difficult and will require innovative NLP approaches.

Acknowledgements This work was supported by the National Library of Medicine (NLM) grant 1K99LM012104, as well as the intramural research program at NLM.

9. Bibliographical References

Andersen, U., Braasch, A., Henriksen, L., Huszka, C., Johannsen, A., Kayser, L., Maegaard, B., Norgaard, O., Schulz, S., and Wedekind, J. (2012). Creation and use of Language Resources in a Question-Answering eHealth System. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2536–2542.

Artzi, Y. and Zettlemoyer, L. S. (2011). Learning compact lexicons for CCG semantic parsing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Artzi, Y., Das, D., and Petrov, S. (2014). Learning Compact Lexicons for CCG Semantic Parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1273–1283.

Artzi, Y., Lee, K., and Zettlemoyer, L. (2015). Broad-coverage CCG Semantic Parsing with AMR. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Athenikos, S. J. and Han, H. (2010). Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99:1–24.

Athenikos, S. J., Han, H., and Brooks, A. D. (2009). A framework of a logic-based question-answering system for the medical domain (LOQAS-Med). In *Proceedings of the 2009 ACM Symposium on Applied Computing*, pages 847–851.

Cairns, B. L., Nielsen, R. D., Masanz, J. J., Martin, J. H., Palmer, M. S., Ward, W. H., and Savova, G. K. (2011). The MiPACQ Clinical Question Answering System. In *Proceedings of the AMIA Annual Symposium*, pages 171–180.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., Ely, J., and Yu, H. (2011). AskHERMES: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics*, 44:277–288.

Clarke, J., Goldwasser, D., Chang, M., and Roth, D. (2010). Driving semantic parsing from the world's response. In *Proceedings of the Conference on Computational Natural Language Learning*.

Demner-Fushman, D. and Lin, J. (2007). Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1):63–103.

Hanauer, D. A., Mei, Q., Law, J., Khanna, R., and Zheng, K. (2014). Supporting information retrieval from electronic health records: A report of University of Michigans nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *Journal of Biomedical Informatics*, 55:290–300.

Jimeno-Yepes, A., McInnes, B. T., and Aronson, A. R. (2011). Collocation analysis for UMLS knowledge-based word disambiguation. *BMC Bioinformatics*, 12(Suppl 3):S4.

Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. (2013). Interpreting Consumer Health Questions: The Role of Anaphora and Ellipsis. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 54–62.

Kobayashi, T. and Shyu, C.-R. (2006). Representing Clinical Questions by Semantic Type for Better Classification. In *Proceedings of the AMIA Annual Symposium*.

Li, M. (2012). *Investigation, Design and Implementation of a Clinical Question Answering System*. Ph.D. thesis, University of Sydney.

Liang, P., Jordan, M. I., and Klein, D. (2011). Learning Dependency-Based Compositional Semantics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 590–599.

Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993). The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291.

Liu, F., Antieau, L. D., and Yu, H. (2011). Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain. *Journal of Biomedical Informatics*, 44(6).

Muresan, S. (2011). Learning for deep language understanding. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Neumann, L. (2010). Comparative Domestic and International EHR Adoption. In *National Electronic Health Records Summit*.

Office of the National Coordinator for Health Information Technology. (2014). Federal Health IT Strategic Plan: 2015–2020. Available at: https://www.healthit.gov/sites/default/files/federal-healthIT-strategic-plan-2014.pdf.

Patrick, J. and Li, M. (2012). An ontology for clinical questions about the contents of patient notes. *Journal of Biomedical Informatics*, 45:292–306.

Poon, H. (2013). Grounded unsupervised semantic parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W. W., and Savova, G. (2015). Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154.

Roberts, K. and Demner-Fushman, D. (2016). Interactive use of online health resources: A comparison of consumer and professional questions. *Journal of the American Medical Informatics Association*.

Roberts, K., Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. (2014a). Decomposing Consumer Health Questions. In *Proceedings of the 2014 BioNLP Workshop*, pages 29–37.

Roberts, K., Masterton, K., Fiszman, M., Kilicoglu, H., and Demner-Fushman, D. (2014b). Annotating Question Decomposition on Complex Medical Questions. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2598–2602.

Roberts, K., Masterton, K., Fiszman, M., Kilicoglu, H., and Demner-Fushman, D. (2014c). Annotating Question Types for Consumer Health Questions. In *Proceedings of the Fourth LREC Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*.

Roberts, K., Simpson, M. S., Voorhees, E., and Hersh, W. (2015). Overview of the TREC 2015 Clinical Decision Support Track. In *Proceedings of the 2015 Text Retrieval Conference*.

Roberts, K., Simpson, M. S., Demner-Fushman, D., Voorhees, E., and Hersh, W. R. (2016). State-of-the-art in biomedical literature retrieval for clinical cases: A survey of the TREC 2014 CDS Track. *Information Retrieval Journal*, 19(1).

Schardt, C., Adams, M. B., Owens, T., Keitz, S., and Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics & Decision Making*, 7(16).

Schnall, J. G. and Fowler, S. (2013). MedlinePlus.gov: Quality Health Information for Your Patients. *American Journal of Nursing*, 113(9):64–65.

Simpson, M. S., Voorhees, E., and Hersh, W. (2014). Overview of the TREC 2014 Clinical Decision Support Track. In *Proceedings of the 2014 Text Retrieval Conference*.

Stearns, M. Q., Price, C., Spackman, K. A., and Yang, A. Y. (2001). SNOMED Clinical Terms: Overview of the Development Process and Project Status. In *Proceedings of the AMIA Annual Symposium*, pages 662–666.

Terol, R. M., Martinez-Barco, P., and Palomar, M. (2007). A knowledge based method for the medical question answering problem. *Computers in Biology and Medicine*, 37(10):1511–1521.

Van Der Volgen, J., Harris, B. R., and Demner-Fushman, D. (2013). Analysis of Consumer Health Questions for Development of Question-Answering Technology. In *Pro-*

ceedings of the 2013 Annual Meeting and Exhibition of the Medical Library Association.

Voorhees, E. M. and Hersh, W. (2012). Overview of the TREC 2012 Medical Records Track. In *Proceedings of the 11th Text REtrieval Conference*.

Voorhees, E. M. and Tong, R. M. (2011). Overview of the TREC 2011 Medical Records Track. In *Proceedings of the 10th Text REtrieval Conference*.

Waldinger, R., Bobrow, D. G., Condoravdi, C., Das, A., and Richardson, K. (2011). Accessing Structured Health Information through English Queries and Automatic Deduction. In *Proceedings of the AAAI 2011 Spring Symposium*, pages 70–73.

Wong, Y. and Mooney, R. J. (2007). Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Yu, H. and Cao, Y. (2008). Automatically Extracting Information Needs from Ad Hoc Clinical Questions. In *Proceedings of the AMIA Annual Symposium*, pages 96–100.

Yu, H. and Sable, C. (2005). Being *Erlang Shen*: Identifying Answerable Questions. In *IJCAI Workshop on Knowledge and Reasoning for Answering Questions*.

Zettlemoyer, L. and Collins, M. (2005). Learning to Map Sentences to Logical Forms: Structured Classification with Probabilistic Categorical Grammars. In *Conference on Uncertainty in Artificial Intelligence*.

Zhang, J. and Walji, M., editors. (2014). *Better EHR: Usability, workflow, & cognitive support in electronic health records*. National Center for Cognitive Informatics & Decision Making in Healthcare.

Zhang, Y. (2010). Contextualizing Consumer Health Information Searching: An Analysis of Questions in a Social Q&A Community. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 210–219.