Syntactic-Semantic Question Frames for Cohort Identification

Dina Demner-Fushman, MD, PhD, Swapna Abhyankar, MD
National Library of Medicine, Bethesda, MD

Abstract

Large sets of electronic health record (EHR) data are increasingly used in retrospective clinical studies and comparative effectiveness research. Free text is often used to describe the desired patient cohort characteristics for such studies. We present a syntactic-semantic approach to capturing free-text cohort characteristics in a structured frame format. We generated 60 topics to develop the approach, and evaluated it on 30 IOM priority topics for comparative effectiveness research that were provided for the Medical Records evaluation at the 2011 Text Retrieval Conference. We evaluated the accuracy of the frames as well as the modifications needed to achieve near perfect precision in identifying the top 10 eligible patients. Our automatic approach accurately captured 29 test questions, of which 21 needed no modification for finding eligible patients. Overall, the syntactic–semantic frames compared favorably to keyword searches when a domain-specific search engine was used for cohort selection.

Introduction

Cohort identification is an essential phase of clinical research and an active area of medical informatics research. In secondary use datasets, many cohort characteristics can be found only in free-text reports. We propose a two-stage question-answering approach to cohort selection using natural language. In this work, we focus on capturing and formally representing the patient characteristics expressed in free-text requests. We present the automatic generation of question frames for retrieval of relevant clinical reports.

Methods

Building on the evidence-based medicine four-slot **P**atient/**P**roblem, **I**ntervention, **C**omparison, **O**utcome (PICO) well-formed question frame, we developed frames for capturing nuances of the question, such as temporal relations, with relational slots that express question elements using predicate-argument structures ([concept]–(relation)–[concept]). We first developed the ideal slots for the expanded question frame manually using 60 training topics. We then developed an algorithm that uses MetaMap and Stanford parser to automatically extract the frames from the topics. We evaluated the frame extraction algorithm on 30 IOM-based test topics. In addition, we evaluated the quality of cohort identification achieved by the automatically generated frames.

Results

Of the 30 test frames, 29 were accurate and one was incorrect. The reason for the one incorrect frame was the lack of an appropriate cue in our pattern set. After the cue was added to the patterns, the algorithm generated 30 correct frames. When we evaluated the usefulness of the automatic frames for cohort identification, we found that nine out of thirty (30%) needed modification. Of those, six would have failed to find the most relevant patients. The remaining three needed modifications to improve recall and precision. In all cases, modifications required domain knowledge beyond the UMLS synonymy. For the questions that would have failed, the drug classes or high level descriptions of the procedures needed to be expanded with specific instances.

Discussion

Formally representing the essence of the cohort characteristics for comparative effectiveness studies can potentially streamline the cohort identification process. Our primary concern in developing the frames was determining the minimum set of slots needed to capture all necessary fine-grained details. Future work on a broader set of questions for different retrospective studies will determine if our current set of 21 slots is capable of capturing such details. Our pilot evaluation of capturing the free-text inclusion/exclusion criteria of a cost-effectiveness study cohort in a structured form shows that syntactic-semantic frames can accurately capture the desirable patient characteristics.