Knowledge Extraction of Cohort Characteristics in Research Publications

Jay D. S. Franklin[1], Shruthi Chari[1], Morgan A. Foreman[2], Oshani Seneviratne, PhD[1], Daniel M. Gruen, PhD[2], James P. McCusker, PhD[1], Amar K. Das, MD, PhD[2], Deborah L. McGuinness, PhD[1]
[1]Rensselaer Polytechnic Institute, Troy, NY; [2]IBM Research, Cambridge, MA

Introduction

Healthcare providers may wish to assess how relevant a clinical study result is for a specific patient based on how closely the patient's characteristics (e.g., demographics or laboratory results) match those of the study's cohort. We have previously built the Study Cohort Ontology (SCO) to model population characteristics and encode this knowledge as Resource Description Framework (RDF) Knowledge Graphs (KGs), intended for use as a knowledge source in physician-facing applications that enable clinical decision support [1]. In our current work, we seek to populate the SCO KG by automatically extracting the study cohort information from tables in a research publication. Study cohort tables exhibit wide variance in representation, style and content and thus create challenges for direct translation into KGs. We are developing a general, scalable knowledge extraction pipeline to address these challenges. In this paper we describe our methodology for the pipeline, present initial results based on a set of research publications for Type-2 Diabetes and discuss future work to achieve complete automation of KG creation from study cohort tables.

Methods

We use a pipelined approach to gradually identify the structure of study cohort tables and build a preliminary KG. While study cohort table formats vary, certain patterns are widely observed, and each step of the knowledge extraction pipeline uses different heuristics to exploit these patterns.

In Step 1, we start with a PDF file of a research publication. After manually identifying the study cohort table in this PDF, we use IBM's internal Corpus Conversion Service [2] to create a JSON file of the extracted text, font, style, and pixel bounding boxes for each cell contained in these identified tables. The tool is suited to parsing many varying tables, and can identify nested column headers–but not nested rows, which we develop a heuristic to address in Step 2.

In Step 2, the row subheadings from the tables are identified. Our heuristic measures pixel differences between different rows to identify indentation, and infers the nesting of rows. By identifying row parents and children, we reorganize the flat tabular structure into a tree table structure.

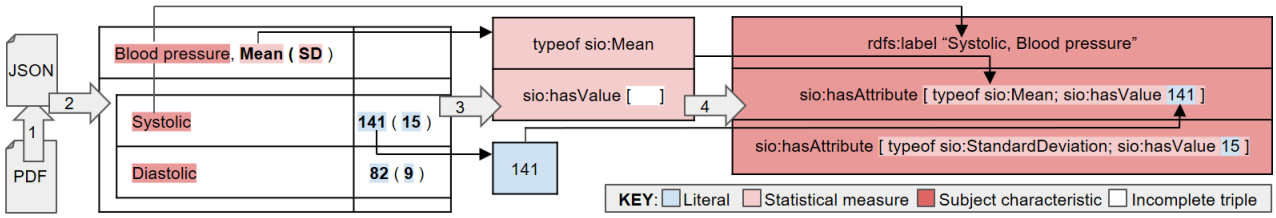

Figure 1: The knowledge extraction pipeline consists of four steps. 1: PDF-JSON conversion, 2: Creation of tree table, 3: Annotating tokens with KG components, and 4: Filling of incomplete RDF triples where possible.

In Step 3, KG components are identified from the text of table cells. The cell's text is first tokenized, according to a regular expression that identifies numbers, punctuation, and alphanumerical words as separate tokens. Each row and column gets their own annotation, i.e., 'subject characteristic' for rows and 'study arm' for columns. For each ontology concept we intend to identify, we maintain a list of recurring terms or keywords that indicate this concept is being measured in a study cohort table. Tokens that match a keyword are annotated with the matching concept. Numerical tokens are annotated with their parsed value. Tokens without a match are labeled with their corresponding row/column header (e.g.in Figure 1, a subject characteristic is labeled 'Systolic, Blood Pressure'). Each of these annotations corresponds to a concept in the provided ontology, and is translatable to an RDF graph component instantiating this concept. However, most of these components are initially incomplete, e.g., a 'mean' node is associated with both a complete triple ('type-of Mean') and an incomplete triple ('has-value unknown-literal-value').

In Step 4, these components (serving as KG nodes) are combined with one another to construct the KG, using the relationships between concepts from the provided ontology. As seen in Figure 1, a mean node could have its 'unknown-literal-value' component filled with a literal value 141, creating 'mean has-value 141'. To form edges between the correct nodes, we perform a depth-first search of the nested tree table structure that we constructed in Step 1. In this way, nodes from row subheadings (such as 'mean' and 'SD' in Figure 1) are combined with nodes in child rows.

Results

To evaluate our pipeline, we selected 18 research publications cited in the pharmaceutical interventions and hypertensive comorbodities chapters of the American Diabetes Association (ADA) Standards of Medical Care guideline 2018 [3]. We only evaluate the efficiency of our pipeline in organizing the tabular content and correctly recognizing it, and thus excluded 2 tables that failed to convert during the PDF-JSON conversion process. Aggregatively, the study cohort tables we considered included 1492 statistical measures recorded for 720 subject characteristics reported on 44 distinct study arms. We use the previously developed SCO ontology [1] for the output KG. The SCO ontology reuses concepts from standard, well-used biomedical ontologies, to support the modeling of study cohort table components, and enables a system that supports cohort similarity applications and population analysis scenarios.
Accuracy of parsing the correct numerical value is 97.5%, accuracy of assigning values to the correct statistical measure in SCO is 88.9%, accuracy of assigning statistical measures to the correct subject characteristic grouping is 95.0%. The accuracy of associating a value with its correct statistical measure, subject characteristic, and study arm is 84.0%. Errors are primarily localized in the first two steps of the pipeline, when tabular data is extracted from PDFs and when this tabular data is converted into a nested tree table structure.

Discussion

Currently, our pipeline is able to take a research publication in PDF form, and produce a preliminary KG that assembles the tabular components as per the relationships in the provided ontology. Although this preliminary KG only matches terms to specific concepts when keyword mappings are provided, unidentified terms are included as metadata associated with placeholder concepts. Prior work [4] shows that keyword-matching can be used to identify tabular data from clinical literature, but requires specific rules for each extracted data type (e.g. characteristics such as BMI or age). Our pipeline avoids this issue of manually designing or training these rules by incorporating relationships between data types, already described by existing ontologies, into the extraction pipeline. Our initial results show that we are able to mitigate variance in the format of study cohort tables, as we have been able to identify the statistical measures of subject characteristics, and build a KG encapsulating associations between these measures and other components of the table. We can improve our results by improving the heuristics used for identifying row sub-headers, as some tables do not use indentation to indicate these. In order to make the system scalable, we plan to augment the manually assembled keyword mappings with methods based on Unified Medical Language System (UMLS) tagging, such as UMLS MetaMap [5]. Overall, the KGs we are creating show the validity of an ontological approach to extracting study cohort data from tables and are a step in the automatic recovery of clinical trial data for analysis purposes.

References

1. Chari S, Qi M, Agu NN, Seneviratne O, McCusker JP, Bennett KP, et al. Making study populations visible through knowledge graphs. In: International Semantic Web Conference (ISWC). Auckland, New Zealand; 2019. p. 53–68.
2. Staar PW, Dolfi M, Auer C, Bekas C. Corpus conversion service: A machine learning platform to ingest documents at scale. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM; 2018. p. 774–782.
3. Riddle MC. Standards of medical care in diabetes. Diabetes Care. 2018;41(1). Available from: https://diabetesed.net/wp-content/uploads/2017/12/2018-ADA-Standards-of-Care.pdf.
4. Milosevic N, Gregson C, Hernandez R, Nenadic G. Extracting patient data from tables in clinical literature-case study on extraction of BMI, weight and number of patients. In: HEALTHINF; 2016. p. 223–228.
5. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010 05;17(3):229–236. Available from: https://doi.org/10.1136/jamia.2009.002733.