

Schema Builder: A Web-based User Interface for Authoring and Sharing Natural-Language Processing Schemas

Yang Liu¹, Melissa Castine¹, Matthew Hong¹, Harry Hochheiser PhD²,
Wendy W. Chapman PhD¹

¹University of California, CA; ² University of Pittsburgh, PA

Abstract

We present *Schema Builder*, a web-based user interface to create NLP schemas representing variables to be extracted from clinical text. *Schema Builder* allows the user to create rich concept-oriented, unambiguous schemas that consist of definitions, linguistic and semantic attributes, synonyms, and mappings to standardized vocabularies. The application assists the user by leveraging a detailed generic NLP schema and providing candidate lists for synonyms, definitions, and concepts from standardized vocabularies. Schemas can be shared among users.

Introduction

Clinical research, quality assessment, and decision support often require targeted extraction of information from clinical narratives. Natural Language Processing (NLP) systems generate multi-layered, complex annotations that can be queried to assess the presence of the targeted information; however, extensive NLP expertise is required to create the queries. *Schema Builder* addresses the gap between the NLP system output and the targeted information needs of domain experts by providing a user interface to assist domain experts in building rich concept-oriented, unambiguous schemas that can be used to map to NLP annotations. It will also facilitate collaborative development.

Schema Builder Use Case

A researcher wishes to assess the quality of colonoscopy exams. He *logs in* to *Schema Builder* and *creates a schema* for the project. First, he *creates a patient variable* to identify patients **at least 50 years of age**. *Schema Builder* leverages our generic schema ontology to structure the variable based on the patient class and its allowable properties (e.g., age, race, sex). He *creates a disease variable* for **colorectal cancer** and selects linguistic modifiers for certainty, experiencer, and temporality to expand the variable to **family history of colorectal cancer**. *Schema Builder* queries the UMLS and presents concept candidates. The researcher *selects the best matching concept* and can apply the definition provided by the UMLS, Wikipedia, or other online sources. Second, the researcher can assign values to attributes that are specific to the semantic type of the variable. For instance, he created a variable for the *medication* **Moviprep bowel preparation** and can therefore specify the dosage, route, and frequency. Third, he *checks multiple synonym candidates* that are retrieved from the UMLS, the Semantic Medline Database, and a large clinical report database. A new researcher creates a disease variable for colorectal cancer and finds there is already one created by a previous user. He *clones this existing variable* to his account and *adds a linguistic attribute* “Negated”, resulting in a variable **No family history of colorectal cancer** (as shown in the figure). The output of *Schema Builder* is an RDF or OWL file of NLP schemas instantiated with variables and their attributes that the user wants to identify in the text. The variables in the schema can now be used to query output from an NLP system to determine whether the variable occurs in a report.

System Design

Schema Builder is a multi-user web application built on Grails with rich HTML5 interactions provided by Dojo. Each user is permitted to create, edit and share schemas, and clone shared schemas or variables from other users. We use MongoDB as a document-oriented database to store schemas for high performance and also as a memory cache for fast searching.

Future Work

We are working on creating relationships between variables. After that, we will carry out user studies and perform iterative user interface design. We will use the output of *Schema Builder* to retrieve patient data and display relevant NLP annotations for retrospective chart review.

Acknowledgment: This study is funded by NIH 1R01LM010964, VAHSR&D HIR 08-204.

Definition Attributes Synonyms

Variable Name: No Family History of Colorectal Cancer

No Family History of Colorectal Cancer

Search in Knowledge Base

Link to Reference

Link	Name	Source	Categories
<input type="checkbox"/>	Colorectal Carcinoma	UMLS	Neoplastic Process
<input checked="" type="checkbox"/>	Colorectal Cancer	UMLS	Neoplastic Process
<input type="checkbox"/>	Colorectal Cancer Pathway	UMLS	Functional Concept