Name:	Noémie Elhadad
Affiliation:	Columbia University
Email:	noemie@gmail.com
Presentation type (select one):	Poster

Leveraging Clinical Texts and Enabling Natural Language Processing in OHDSI

OHDSI Natural Language Processing Working Group

Abstract

Clinical narratives and template notes contain a wealth of information about patients' history, health status, socio-economic history, and the healthcare processes that guide their care. While Natural Language Processing is still an active area of research, open-source NLP tools have matured enough to be incorporated in many applications and tasks relevant to the OHDSI mission. In this poster, we report on the ongoing efforts of the OHDSI NLP Working Group to incorporate the output of NLP tools into the OHDSI CDM.

Introduction

The mission of the OHDSI Natural Language Processing (NLP) Working Group is to promote the Promote the use of textual information from EHRs for observational studies under the OHDSI umbrella. In practice, the working group has been working towards (1) a schema to enable storing NLP output of clinical texts in the OHDSI CDM; (2) disseminating to the OHDSI community examples IRBs for use of clinical texts; (3) streamlining open-source NLP tools for parsing of clinical texts and software pipelines for ETL; and (4) collecting use cases and studies to showcase on the value of NLP in the OHDSI ecosystem. In this poster we present our ongoing efforts towards the first aim. The poster will motivate some use cases, our proposed edits to the existing Note table in the CDM, and our proposed new Note_NLP table to store the output of NLP software. Finally, we will discuss the potential future extensions to the information model and the use of more efficient storage for textual and NLP output.

Use Cases

With the goal of a pragmatic and practical NLP schema that answers the needs of current tasks relevant to the OHDSI mission, the NLP working group collectively discussed use cases, including implementing phenotyping algorithms at scale [1], extracting clinically relevant measures that are conveyed in texts, such as ejection fraction [2], cohort selection [3], survival analysis for disease progression [4], and patient-level visualizations [5].

Note Table

The OHDSI CDM already contains a Note table. The NLP working group proposes a few edits to generalize its use across OHDSI members (e.g., a field to encode language of the note) and to standardize the descriptors of the note to LOINC-based descriptions according to meaningful axes (e.g., clinical setting, clinical domain).

NLP Concepts, Modifiers, and Types

Below is the current proposed new Note_NLP table for inclusion in the CDM. It enables to store data provenance information (NLP_system, NLP_date fields), the extracted concepts and their links to terminology, as well as the most common modifiers most commonly used in our use cases (existence, optional value, and temporal).

Relational Databases and Other Storage for Text and NLP Output

While the current relational database schema makes sense for the Note table, the NLP working group continues to discuss the value of other storage solutions for text and the output of NLP software. Some institutions rely on Lucene indexes, which encode in a compact and query-efficient fashion text along with layers of annotations like the ones provided by an NLP software. In particular, for data-driven models, this solution might prove more practical than storing all NLP output in a relational format.

Conclusion

NLP is mature enough to incorporate its output in the OHDSI ecosystem. The NLP working group is actively working towards facilitating the use of NLP for OHDSI members.

Note_NLP_id	Unique identifier for each concept extracted from NLP
note_id	Foreign key identifier to the note the concept was extracted from (Note table).
section_concept_id	Foreign key to predefined concept identifier in the Standardized Vocabularies (LOINC) reflecting the section the extracted concept belongs to.
snippet	Small window of text surrounding term mention
lexical_variant	Raw text extracted from NLP
Note_NLP_concept_id	Foreign key to concept id (Concept Table). Domain concept is provided as part of the Concept table.
NLP_system	String describing system and version used for NLP (data provenance)
NLP_date	Date describing date at which note was processed
Term_exists	Optional boolean; summary modifier that signifies presence or absence of a term for given patient (e.g., not negated, not conditional, not generic, not uncertain → termmention_ispresent=YES)
Value_as_concept_id	Optional foreign key to standard terminology (e.g., "high"); value of term
Value_as_number	Optional float; potential value of term
Unit_concept_id	Optional foreign key to unit concepts (e.g., "mg/ml"); unit of term value
Term_temporal	Optional string time expression extracted associated to term, "past", "present"
Term_modifiers	Optional string describing pipe-delimited other modifiers (e.g., course, severity, etc.) extracted by NLP software.

References

1. Pivovarov R, Perotte A, Grave E, Angiolillo J, Wiggins C, Elhadad N. (2015) Learning probabilistic phenotypes from heterogeneous EHR data. *J Biomed Inform.* 58:156-165.
2. Patterson OV, Freiberg MS, Brandt C, DuVall SL. Unlocking echocardiogram report measures for heart disease research through natural language processing. In preparation
3. Duke JD, Chase M, Ring N, Martin J, Fuhr R, Hirch A. (2016) Natural Language Processing to Augment Identification of Peripheral Arterial Disease Patients in Observational Research. *American College of Cardiology Annual Symposium*.
4. Perotte A, Ranganath R, Hirsch J, Blei D, Elhadad N (2015). Risk Prediction for Chronic Kidney Disease Progression Using Heterogeneous Electronic Health Record Data and Time Series Analysis. *J Am Med Inform Assoc*. 22(4):8720
5. Hirsch J, Tanenbaum J, Lipsky Gorman S, Liu C, Schmitz E, Hashorva D, Ervits A, Vawdrey D, Sturm M, Elhadad N. (2015) HARVEST, a longitudinal patient record summarizer. *J Am Med Inform Assoc*. 22(2):263-274.