Sarah MacHarg
Fall 2020 Civic Digital Fellow - Data Science - NIDDK
Supervisor: Kenneth Wilkins

Developing an NLP Pipeline to Tag and Extract CDEs from NIDDK Repository Documents

The NIH Common Data Element (CDE) Repository maintains a collection of 40,000+ CDEs in machine-readable formats. The general motivation behind this project was to standardize and promote the use of these common data elements, and the metadata to distinguish them from other data elements in use, with the goal of making data collected by NIH-funded studies more FAIR (Findable, Accessible, Interoperable, and Reusable). This is especially important given the context of COVID-19, as NIDDK-funded studies contain swaths of potentially reusable data on obesity, diabetes, and social determinants of health, which are all emerging as key COVID risk factors. Furthermore, the NIDDK Central Repository (DKCR) is just one of many NIH-maintained repositories that manage resources from multiple multi-center and large single-center clinical studies. Given this landscape, the goal of this project was to improve researchers' ability to harness this wealth of data through two key objectives:

1. Develop a Natural Language Processing (NLP) pipeline that detects CDEs in repository studies
2. Draft recommendations for improving repositories' metadata collection and access procedures

Developing an NLP Pipeline

I first interviewed stakeholders and identified the following key features of common data elements:

- Uniqueness
- Clearly defined permissible values
- Context
- Variable description
- Units of measurement
- Datatype
- Data Parameters (e.g. max, min)
- Measurement protocol

I then developed custom functions to parse these features, if present, from CDEs and DKCR datasets. Because metadata formats were inconsistent in both the CDE and NIDDK repositories, this step was more time-consuming than expected, albeit very informative in drafting recommendations for repository metadata curation. After preprocessing, I explored two match discovery methods: probabilistic record linkage with the Fellegi-Sunter model, and fuzzy string matching with the `fuzzywuzzy` package. For the former, I used `scispacy`'s entity detection capabilities to detect UMLS Metathesaurus Concept Unique Identifiers (CUIs) in variable text. I represented those CUIs as numeric vectors using Harvard's `cui2vec` embeddings, then compiled vectors into a matrix to represent a full data element. After creating custom methods to compare vectors of data element features that included CUI matrices, datatypes and parameters, I ran the record linkage algorithm. Unfortunately, this model drastically underperformed, selecting thousands of false matches in a way that seemed no better than choosing by random chance. This likely was due to the high proportion of potentially false matches in the set of candidate pairs, and prompted me to explore `fuzzywuzzy` as an alternative method for data element comparison.

Using `fuzzywuzzy` to compare variable and permissible value descriptions yielded far more promising results. While the pipeline is unable to identify data element-to-CDE matches with high accuracy, we made significant steps in wrangling the data, exploring the solution space and automatically narrowing down the set of all CDEs to a few likely matches per data element. These candidate sets consistently include high quality matches. An attempt at using record linkage after `fuzzywuzzy` to eliminate some of

the remaining false matches yielded results little better than chance, at best. Thus, future work will need to explore other fine-tuning methods. Labelling a dataset with matching CDEs, which requires many domain-expert hours of feature engineering, would also allow for *quantitative* analysis of model performance, which will be useful in the model-refining process.

Repository Recommendations and Future Steps

Within the NIH CDE Repository, a key obstacle was lack of standardization. Data elements are inconsistently recorded—for example, some CDEs specifically designate the units of measurement in the proper field, while others simply include it in the free text description of the variable, adding challenges and ambiguity to any CDE-related work. Duplicate and near-duplicate CDEs also seem common. Both these phenomena make using CDEs more difficult for those prospectively planning studies and those integrating data for re-use (whether trained as biomedical researchers or data scientists outside biomedicine). Inconsistencies in preexisting CDEs could be addressed with other NLP pipelines in the future, informed by findings from this project, while formulating a set of strict and explicit guidelines for new CDE submissions or 'tagging' of existing NIH CDE Repository entries would prevent further inconsistencies. As for apparent duplication, the code for my pipeline can easily be adjusted to compare and detect similar CDEs. Once pipeline performance is improved, this would facilitate the apparent-duplicate-entry curation process, among others.

Within the DKCR, inconsistency of metadata structure across studies makes it incredibly difficult to identify which studies are relevant to one's research, find variables within those studies, or parse study data. This poses an immediate obstacle to promoting Findability as part of the NIH goal of FAIR data, let alone the goals for data re-use that led to the DKCR's initiation in 2003. It took weeks to identify, locate and access the metadata I needed. This required level of effort demonstrates how much data-contributors' practices can diverge when not held to a standard. It also clearly implies that, in order to meet the goal of having FAIR data, NIH-hosted data repositories need to devote resources for post-hoc curation and annotation of existing data sets. This points to actionable follow-up by any NIH Institute, Center, or Office (ICO) looking to support the NIH Strategic Plan for Data Science and its Final Policy for Data Management and Sharing. Developing standards for metadata submission that grantees are required to comply with as a condition of their funding would significantly streamline data analysis and reuse. From there, using a refined version of my pipeline to tag CDEs would further increase FAIRness. I have drafted additional documents describing issues I encountered with the repository and potential solutions in more detail. In the new year, I will be presenting some of these experiences and ideas to the CDE Task Force to help spark a cross-ICO discussion on repository requirements.

Looking longer term at the future of this project, once the pipeline is refined, the initial parsing capabilities could be expanded to include a variety of other resources, such as grant applications and contract proposals. As CDE curation improves, this tool could be key in tracking and promoting CDE uptake across the research community.