INDEX: the Intelligent Data Steward Toolbox

Utilizing Large Language Model Embeddings for Automated Data Harmonization

Tim Adams[1], Mohamed Aborageh[1], Yasamin Salimi[1,2], Holger Fröhlich[1,2] and Marc Jacobs[1]

[1]*Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin 53757, Germany*
[2]*Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany*

Abstract

The data steward, responsible for overseeing data management, plays a pivotal role in evidence-based medicine by ensuring the quality, integrity, and accessibility of data throughout its lifecycle. However, managing medical data poses challenges, including handling diverse structured and unstructured data from various sources in different formats. This data curation process demands significant time and resources. To alleviate these challenges and enhance the efficiency of data stewards, we introduce a novel data stewardship tool and curation workflow utilizing Large Language Models (LLMs). We evaluated our approach by performing automatic pairwise cohort harmonization using data dictionaries of 6 different Parkinson's Disease (PD) studies and 13 different studies in the context of Alzheimer's Disease (AD), as well as a mapping task of over 38,000 ICD10 codes using code descriptions obtained from UKBioBank. When compared with a String Matching based baseline method that does not capture the context of variable descriptions, we found that Generative Pre-trained Transformer (GPT) embedding based mappings performed significantly better, reaching a best average accuracy for the application of PD cohort harmonization for an automated initial closest match of 82%. While we found that due to various different formulation and wording issues descriptions could not be automatically matched in all cases, we are confident that our data steward tool can significantly facilitate the work of the data steward in a semi-automatic fashion.

Keywords

data stewardship, large language models, embeddings, semantic mappings, common data model

As data stewardship is an important but often time and resources intensive process, data stewardship tools can be used to facilitate the process effectively. Variable descriptions for data harmonization are often very diverse in their formulation; it is therefore important to incorporate their semantics to be able to harmonize them with a high accuracy. With the ongoing development of GPT models, we evaluated whether vector distances of GPT model embeddings can be used to automatically harmonize variable descriptions. We developed a data steward tool and a harmonization workflow [1] that can be used to iteratively improve harmonization results in a semi-automated process.

We evaluated our automated mapping approach based on three different application cases: We harmonized 6 different Parkinson's Disease (PD) cohorts pairwise using GPT- embedding

SWAT4HCLS'24: Semantic Web Applications and Tools for Health Care and Life Sciences, Feb 26–29, 2024, Leiden, NL
tim.adams@scai.fraunhofer.de (T. Adams); mohamed.aborageh@scai.fraunhofer.de (M. Aborageh); yasamin.salimi@scai.fraunhofer.de (Y. Salimi); holger.froehlich@scai.fraunhofer.de (H. Fröhlich); marc.jacobs@scai.fraunhofer.de (M. Jacobs)


CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://github.com/SCAI-BIO/index & https://github.com/SCAI-BIO/index/tree/main/doc/workflow

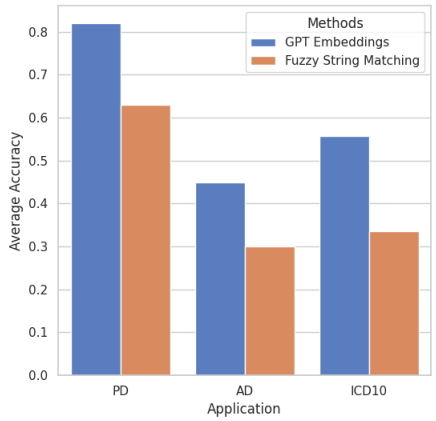

Figure 1: Average accuracy for the three evaluated harmonization tasks.

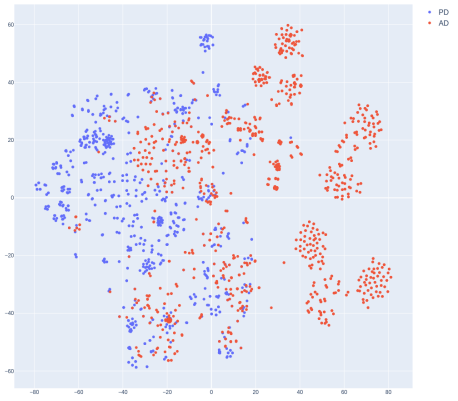

Figure 2: Two-dimensional t-SNE representation of computed AD and PD embeddings.

and Fuzzy String Matching as a baseline comparison, using an in-house Common Data Model (CDM) for ground-truth data. The same was tested in the context of Alzheimer's Disease (AD) using 13 different collected studies. We mapped over 38,000 Read codes for medical diagnosis to ICD10 codes using code descriptions obtained from UK Biobank and referring to a pre-existing mapping as ground truth. Notable examples of correct and incorrect matches are shown in Table1. We tested each approach against a baseline method using Fuzzy String Matching. The results are shown in Figure 1. We found that GPT-Embedding based matching outperformed the baseline method significantly in all three tested application cases, reaching an average accuracy of 82% for the PD cohorts, 63% for the AD mappings and 56% for the automatic mapping of ICD10 codes. Especially for the harmonization application, we found that semantically coherent variable descriptions from different cohorts form distinct clusters that may overlap for different studies, even for different disease types (see Figure2). We however also found that given the very much different ways to formulate data descriptions when taking into account special cases such as custom abbreviations (see Table1), fully automatic data harmonization using LLMs is not yet feasible. We expect that with the ongoing development of LLMs and especially domain trained models, we will be able to further improve and build on our results in the future.

Source Read Description	Matched ICD10 Description	Correct ICD10 Description	Logic
FH: Stomach cancer	Family history of malignant neoplasm of digestive organs	-	True
Cardiac function test abnormal	Abnormal results of cardiovascular function studies	-	True
Macrocytosis	Macroglossia	Other specified diseases of blood and blood-forming organs	False
FH: Depression	Unhappiness	Family history of other mental and behavioral disorders	False

Table 1
Examples of mapped Read and ICD10 descriptions. The "Logic" column indicates a correct match.