Automating the Generation of a Functional Semantic Types Ontology with Foundational Models

Sachin Konan, Scott Affens, Larry Rudolph

Functional Semantic Types (FST)
- Semantic Types are "entity tags" that relate a column to its real-world entity
  - Useful for discovery/search
  - Lack context of data's source, units, and validations
- We introduce Functional Semantic Types (i.e., FST)
  - Types are represented as Python Classes and have relevant functions
  - Necessary for automated discovery/search/normalization/validation/joining

Problem Formulation
- Given a universe of data, extract and generate the FSTs that span all the columns
- To generate FSTs at scale, we use Large Language Models, specifically GPT4

FST and Graph Generation
- Our graph represents a hierarchy of functional generalization, meaning:
  - FSTs at the general layer can normalize more representations of the same entity than those at the table/product layers
- We use LLMs to generate FSTs and edges from G-FST -> F-GST.

Evaluation
- Normalization code had complex behavior, including string normalization, type-casting, external library usage, and more
- The generated code raised runtime exceptions in less than 2% of cases across the universes
- Implies that the code was well-constructed or there were few occurrences of invalid values
- Human evaluation showed that LLMs were performant at entity detection, even without a class distribution
- Additionally, the FSTs were generally well-scaled, but for domain-specific data, they were too general

Applications
- Data Discovery/Search
  - LLMs identified a relationship between two tables referring to life-expectancy, but "life-expectancy" wasn't in either header
- Income is represented differently across tables: string range or float
  - LLM reconciled the differences
  - Validation check asserts correctness
- With transformation logic in functions, it can be easily grepped

Data Fusion
- LLMs can use libraries to perform complex mappings, like currency conversions.

Data Validation
- With wide background knowledge, LLMs can generate meaningful validations
- Precipitation: "T" means trace amounts of water
- Acceleration: Normalized values should be [-1,1]

Disclaimer: The views expressed herein are solely the views of the author(s) and are not necessarily the views of Two Sigma Investments, LP or any of its affiliates. They are not intended to provide, and should not be relied upon for, investment advice.