Unleashing the Potential of Data Lakes with Semantic Enrichment Using Foundation Models

Nandana Mihindukulasooriya[†], Sarthak Dash, Sugato Bagchi, Faisal Chowdhury, Alfio Gliozzo, Ariel Farkash, Michael Glass, Igor Gokhman, Oktie Hassanzadeh, Nhan Pham, Gaetano Rossiello, Boris Rozenberg, Yehoshua Sagron, Dharmashankar Subramanian, Toshihiro Takahashi, Takaaki Tateishi and Long Vu

IBM Research AI

Abstract

Nowadays most organizations are managing data lakes containing heterogeneous data from various sources. However, the lack of adequate metadata often transforms these data lakes into data swamps, making it challenging to locate relevant data for critical organizational tasks and consequently limiting their utility. Recent advancements in large language models and foundation models have enabled the automation of metadata generation using generative AI models and the use of generated metadata for mapping tabular data into semantically richer glossaries, taxonomies, or ontologies.

In this talk, we will present a semantic enrichment process that generates table metadata such as descriptive table captions, tags, expanded column names, and column descriptions and then uses that information to map table columns to concepts in a given business glossary or an ontology. Furthermore, during this process, we represent both table metadata and business glossaries as knowledge graphs and connect them by mapping columns to business concepts. As a result, the enrichment process makes the data in data lakes more meaningful to the organization and enhances downstream tasks, including improved table search and discovery, efficient table joins, and advanced business analytics.

Keywords

Data Lakes, Knowledge Graph, Semantic Enrichment, Large Language Models, Foundation Models

Introduction The use of data lakes by organizations to handle large volumes of structured, semi-structured, and unstructured data from multiple sources is becoming a common practice. Nevertheless, tables in these data lakes often suffer from issues such as abbreviated column names, missing table or column descriptions, tags and other metadata. Lack of adequate metadata in data lakes can limit their usefulness and hinder the relevant data from being found and utilized efficiently in downstream tasks. In this talk, we will present the current challenges of data lakes in an industrial setting and how we are addressing those challenges with a semantic enrichment process using both large language models as well as knowledge graphs.

The current semantic enrichment academic benchmarks often do not sufficiently reflect the challenges in industrial data lakes. The majority of academic benchmarks for semantic enrichment tasks, such as the Column Type Annotation (CTA) task in the Semantic Web

ISWC 2023 Industry Track, November 06–10, 2023, Athens, Greece

[†]Corresponding author.

nandana@ibm.com (N. Mihindukulasooriya)

0000-0003-1707-4842 (N. Mihindukulasooriya)

Challenge on Tabular Data to Knowledge Graph Matching (SemTav), use public domain data that includes entities that can be linked to open knowledge graphs such as DBpedia or Wikidata. With such settings, systems can link the cells to entities in KGs first and then perform enrichment tasks based on those linked entities. Nevertheless, in an enterprise setting, there are a number of additional challenges. First, the table and column names are often abbreviated using data-owner-specific codes or acronyms with minimum or no textual descriptions. This makes it harder to search and discover tables using keyword or semantic search.

Furthermore, most organizations only permit semantic enrichment processes to access to the table metadata such as column headers and not actual data (*i.e.*, cell values) due to privacy and access control regulations. Even when the data is available, they consist of entities that are not present in public knowledge graphs, thus can not be linked. Therefore, in most industrial settings, automatic metadata generation and mapping table columns to concepts using only table metadata become necessary.

Semantic Enrichment Process The inputs to our semantic enrichment process are a set of table metadata (table names and column headers) from a data lake and a business glossary that defines the concepts of interest to the organization. The semantic enrichment process consists of three steps: (a) column name expansion; (b) table metadata enrichment; and (c) column-to-concept mappings (also known as Column Type Annotation or CTA).

The objective of the column name expansion step is to generate meaningful column names for abbreviated and coded cryptic column names using adjacent column names of the same table and table name as the context. The perplexity of language models along with some clues from the business glossaries are used for the step. For table metadata generation, decoder-only style auto-regressive models similar to GPT-4 / Llama 2 or encoder-decoder style sequence-to-sequence models similar to FLAN-T5 are trained either using public open data from portals such as data.gov or industrial data when available. The column-to-concept mapping implementation uses Sentence Transformer (SBERT) models to compute similarities between a column metadata representation and a business glossary term representation. As a result of this process, the tables in the data lakes be annotated with both human-readable descriptions and tags as well as business concepts from glossaries.

Table metadata can be represented as a knowledge graph containing tables and columns with their relationships. Similarly, the glossary concepts can also be represented as a knowledge graph where concepts are linked using relations such as subclass of or part of. Through the process of semantic enrichment, we connect these two knowledge graphs together, unveiling the semantics of the columns and enhancing their utility for downstream tasks.

Conclusion The advancements in large language models have enabled automatic metadata generation for tables using generative AI models and mapping columns to concepts in glossaries using models such as sentence transformers. By representing table metadata, business glossaries, and the mappings between them in a single knowledge graph, downstream applications such as table search and discovery or automatic table joins can utilise this information effectively. In this talk, we plan to discuss the semantic enrichment challenges in an industrial setting, our approach to addressing those challenges, lessons learned, and future directions.