Generating Semantic Context for Data Interoperability in Relational Databases using BGE M3-Embeddings

Bugra Altug, Martin Weise, Andreas Rauber

Abstract

Relational databases store a significant portion of the world’s most valuable data by managing data in tables. It is a common misconception that by providing expressive table- and column names, researchers have sufficient context to reuse the data adequately. Without proper (machine-)understandable context, researchers face a data interoperability problem, i.e. the challenge to confidently interpret data due to lack of context. This problem becomes evident when sharing data in data repositories such as DBRepo [1] that make it findable, accessible, interoperable and reusable to anyone globally.

To aid researchers in the difficult task of mapping a relational database schema to ontologies that describe the conceptual and quantitative context, we use a structure-level matching method based on machine learning. Our semi-automatic mapping system utilizes the BGE M3-Embedding model [2] to encode column names and ontology entity labels. It employs a cosine similarity to identify the best-matching ontology concept for each table column.

This approach successfully matches 89.9% of the contextual correct semantic concepts and units of measurements within the first 10% of all entities (entity coverage) and achieves a Mean Reciprocal Rank (MRR) of 0.5259, outperforming all other approaches. A similar approach is employed to calculate the similarity between columns and units of measurement entities, with an encoding method that adds the "unit" keyword at the end of entity labels. This achieves a 64.4% entity coverage and 0.1164 MRR, also surpassing all other tested approaches.

Let $c_i=\{c_1, \ldots, c_i\}$ with $1 \le i \le n$ be a column of a relational database table schema. Each c_i may optionally have exactly one concept entity (semantic concept) e_{ci} and exactly one unit of measurement entity e_{ui} assigned. Based on the column names $C=\{c_1, \ldots, c_n\}$, our approach suggests top-level ontologies as well as semantic concepts and unit of measurement for each column. Users can correct the suggested semantic concepts by selecting different concept entity mappings. This selection influences the identification of the respective column’s unit of measurement.

We evaluated the efficacy of our method through observing expert users that were trained briefly on the user interface. On average, a researcher needs to click 1.36 times on average

to correctly map a column c_k to a concept entity e_{ck}. Our method is 9.8 times faster than mapping the entity names manually (i.e. typing them). For correctly mapping a unit of measurement entity e_{uk} to a column c_k, our approach requires the researcher to click 5.745 times on average. This makes our method 2.48 times faster than manually typing the entity names. Note that the minimum number of clicks is calculated by simulating user interactions such as clicking on a drop-down from our user interface is counted as one click.

[1] Weise, M., Staudinger, M., Michlits, C., Gergely, E., Stytsenko, K., Ganguly, R., & Rauber, A. (2022). DBRepo: a Semantic Digital Repository for Relational Databases. International Journal of Digital Curation, 17(1), 11. DOI: 10.2218/ijdc.v17i1.825

[2] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation (Version 4). DOI: 10.48550/ARXIV.2402.03216