Schema Matching of Health Data Using Foundation Models

Marcel Parciak [1,2], Brecht Vandevoort [2], Frank Neven [2], Liesbet M. Peeters [1,2] and Stijn Vansummeren [2]

[1] UHasselt, BIOMED, Agoralaan, 3590 Diepenbeek, Belgium
[2] UHasselt, Data Science Institute, Agoralaan, 3590 Diepenbeek, Belgium
{marcel.parciak,brecht.vandevoort,frank.neven,
liesbet.peeters,stijn.vansummeren}@uhasselt.be

Schema matching is a core task in data integration, where the aim is to identify correspondences between schema elements from a source schema and a target schema so that, ultimately, data engineers can map values from source to target. We typically differentiate between name-based, i.e. based on names of schema elements, and instance-based, i.e. based on data instances, matching approaches [1]. Recent work focuses on improving the latter approach [3] and combining it with simple name-based matchers. This is challenging when instances are unavailable, such as in privacy-sensitive environments like the health domain. We tackle this challenge by focusing on name-based schema matching approaches.

In recent years, foundation models have shown great promise for data integration tasks [2]. Following this approach, we study name-based schema matching using large language models (LLMs), namely ChatGPT, using names and descriptions of both attributes and relations to drive the schema mapping process.

In this talk, I will present preliminary results towards answering two questions:

(a) Is an LLM a viable approach for schema mapping in the health domain?

(b) Which limitations does this approach have?

To answer (a), I will present experiments that include testing different prompts, varying amounts of context information and approaches to evaluate answers from the LLM. The evaluation is done based on a publically available ETL[1]. In response to (b), I will present groups of classification errors and other observations during these experiments, highlighting common pitfalls with foundation models for schema matching tasks.

References

[1] AnHai Doan, Alon Halevy, and Zachary G. Ives. *Principles of Data Integration.* Morgan Kaufmann, Waltham, MA, 2012.

[2] Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. Can Foundation Models Wrangle Your Data?, May 2022.

[3] Fahad Ahmed Satti, Musarrat Hussain, Jamil Hussain, Syed Imran Ali, Taqdir Ali, Hafiz Syed Muhammad Bilal, Taechoong Chung, and Sungyoung Lee. Unsupervised Semantic Mapping for Healthcare Data Storage Schema. *IEEE Access*, 9:107267–107278, 2021.

[1] https://github.com/OHDSI/ETL-Synthea/