Benchmarking Tabular Representation Models on Longitudinal Data [Talk]

PRATIK POKHAREL, University at Buffalo, USA
OLIVER KENNEDY, University at Buffalo, USA

In this talk, we will discuss our efforts to develop a data integration solution for social science researchers conducting longitudinal studies. We will present a comparative study of the effectiveness of state-of-the-art open tabular representation models for table union search workload over a longitudinal survey dataset. While classical and longitudinal data integration share fundamental challenges, existing tabular representation models for open datasets implicitly assume that semantics are grounded in cell values and external entities, such as knowledge bases or distributional semantics from pretrained language models. Longitudinal survey datasets violate this assumption: semantics are defined by measurement instruments and their latent constructs, rather than solely by the data they produce. Additionally, the incremental and structured nature of longitudinal studies makes them a unique case for data integration. In particular, attributes are derived from prose questions rather than semantically rich identifiers, and these questions frequently co-refer. We outline an approach that incorporates these characteristics into column representation models.

1 Problem Statement

Longitudinal study schemas are inherently dynamic, evolving over time as questionnaires are added, removed, rephrased, or restructured to meet the changing goals of the study at different points. For example, an early survey may ask whether a respondent has children, while later waves introduce follow-up questions such as "If yes, how many children do you have?" Similarly, questions may be modified or removed due to cultural, social, or religious sensitivity. Because of the fluidity in data collection, each wave constitutes a distinct dataset. Individually, these datasets are heterogeneous, but collectively they represent the same study and can be viewed as semi-homogeneous. To analyze such data, researchers must integrate these multiple dataset versions by establishing correspondences between variables across waves based on semantic similarity.

Example 1.1. A government agency has been performing surveys bienially since 1972 to monitor trends in consumer opinion. The agency has produced 35 datasets, each with around 100 questions; each from one year of the study. Because of minor changes in the questionnaire each year, these datasets follow similar, but not identical schemas, and must be integrated to be used. Classical data integration techniques that independently integrate each dataset into a global schema, result in a need for the curator to validate combinations of 3500 attribute mappings. Further complicating matters, classical attribute alignment tools designed for short, information rich attribute names fail on the more vague prose questions used to identify each attribute of the survey.

While the problem of longitudinal study data integration is related to general data integration problems, it carries its own unique challenges. In this talk, we will present our preliminary work developing a longitudinal data integration benchmark, including our efforts to understand how existing column representation techniques fare when faced with open-ended prose attribute descriptions and literal data instances. The state of the art in table respresentation learning is often applied to the problem of table unionability search. We are presently focused on the problem of columnar alignment, and in this talk we will discuss how a simple Metadata-based approach [**? ?**] compares to more sophisticated methods.

Authors' Contact Information: Pratik Pokharel, University at Buffalo, Buffalo, NY, USA, pratikpo@buffalo.edu; Oliver Kennedy, University at Buffalo, Buffalo, NY, USA, okennedy@buffalo.edu.

Example 1.2. The data analyst at the government agency applies a metadata-based approach, using individual prose questions to generate embedding vectors. However, this approach overlooks the presence of conditional and referencing variables common in longitudinal studies. For instance, ***Q1: Have you ever contributed to a political campaign?, Q2: If yes, how much did you contribute?.*** *Q2 applies only to a subset of respondents(the ones that respond "YES" to Q1), and its meaning also depends on Q1. Thus, a column's semantics are defined not only by its text, but also by the conditions under which it is asked. Consequently, data absence (e.g., skipped responses) can itself carry semantic meaning.*

A related work[**?**] incorporates both column semantics and inter-column relationships to generate columnar representations for a related use case of finding unionable tables in data lake settings. It relies on entity linking to external knowledge bases to infer attribute semantics and relationship semantics. Another work[**?**] uses contrastive learning to train column encoders, capturing contextual semantics through co-occurring columns within tables by leveraging a contrastive multi-column pre-training strategy.

2 Experiments

We evaluated the effectiveness of [**? ? ?**] for the use case of table union search in a manually annotated groundtruth of a longitudinal social study that stems from biennial surveys since 1948. On an average each query table has 50 unionable candidates in the groundtruth. We evaluated the Mean Average Precision at K from the retrieved candidates by each of the models. The comparitive results are shown in figure 1.

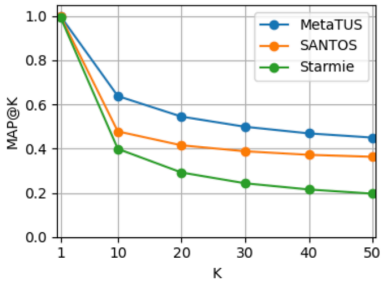

Fig. 1. MAP@K Comparision

We saw a simple "prose question embedding" based model significantly outperform the entity linking based and contrastive learning based model, even when the column name overlap between query tables and candidate tables is signifincantly low (3%). While entity linking based method- SANTOS- is effective for open web tables dominated by named entities, this assumption fails for longitudinal survey data, where values are typically close-ended response literals (e.g., "YES", "NOT VERY", "DEPENDS, PRO-CON"). These literal values are inherently contextual, and neither exist nor should exist in encyclopedic knowledge bases. Moreover, longitudinal survey datasets predominantly use coded, ordinal, and question-specific response labels whose meanings are defined by the survey instrument rather than general language usage. Consequently, the contrastive learning (Starmie) assumption that randomly sampled columns are non-overlapping breaks down in longitudinal settings, where columns across waves are often intentionally overlapping due to construct evolution.

3 Conclusion

Overall, our analysis reveals a modeling mismatch between existing open tabular representation models and the requirements of longitudinal survey data. This mismatch motivates the need for a new benchmark and modeling paradigm tailored to longitudinal studies. Such a benchmark should avoid entity-centric shortcuts and instead require models to reason over survey-specific signals, including question text, response scales, coding schemes, and temporal relationships across survey versions. Correspondingly, future models must incorporate modeling assumptions grounded in survey methodology, explicitly distinguishing between literal surface forms and the latent constructs they measure.