Embedding The Context of Attributes in Longitudinal Study Schemas To Refine Their Semantic Match Candidates

Pratik Pokharel* Oliver Kennedy

University at Buffalo
{ pratikpo, okennedy }@buffalo.edu

Longitudinal study schemas are inherently dynamic, evolving over time. Questionnaires may be added, removed, rephrased, or restructured to meet the evolving demands of the study at different points. For instance, in a long-term investigation of social factors, early surveys might only inquire if the respondent has children. As the study progresses, the need for more detailed information arises, prompting the addition of a follow-up question: "If yes, how many children do you have?". Similarly, some questions may be deemed culturally, socially, or religiously sensitive, or even taboo for certain populations. In these situations, researchers may need to rephrase or eliminate such questions. Because of the fluidity in data collection, each iteration of the survey represents a distinct dataset.

In this talk, we will discuss our efforts to develop a data integration solution for social science researchers conducting longitudinal studies. Although there is significant overlap with classical data integration, the ongoing, incremental, and structured nature of longitudinal studies make them an interesting challenge for data integration [1]. For example, a key differentiating feature between classical data integration and longitudinal studies is that attributes in longitudinal studies result from structured prose questions, rather than being assigned a more semantically dense attribute identifier.

In this talk, we will discuss our work on data integration for longitudinal studies, specifically focusing on the hierarchical structure of longitudinal study survey forms. Specifically, we observe that, while attributes are described in prose (i.e., via questions), this prose is often organized hierarchically and references nearby properties. For example, consider one survey form that includes a category "Family Details" with the question "List out the names and ages of your children", and another that includes a category called "Family," with the questions "How many?" and "What are their ages?" The latter question on the latter form is ambiguous in isolation, but can be related to the first through the category. Analogously, the former question on the latter form is even more ambiguous: (i.e., "How many?" could also be used in a question about raising cattle). The other questions in the category provide context that must be shared for proper integration.

In prior work [1], we leveraged word embeddings to relate questions, but simply concatenated category information into the question's prose. Our talk will explore the use of global as well as the local (neighboring) contextual information of a column to resolve ambiguities and enhance the quality of semantic match candidate suggestions. This approach is not new, for example Zhang [2] exploits signals from the table context, and column values through a deep learning model that predicts semantic types of table columns. However, this approach is not appropriate for us: (i) The semantic type labels in available corpuses like WordNet and DBpedia are too narrow to resolve ambiguity (e.g., "How many?"); (ii) Our use case has a much richer prose description of each attribute; (iii) Our users may lack direct access to column data (which is often heavily controlled due to personally identifying information). We will outline our approach to incorporating embeddings for the category and adjacent attributes into the inter-attribute distance measure used for attribute matching.

References

[1] Pratik Pokharel, Juseung Lee, Oliver Kennedy, Marianthi Markatou, Andrew Talal, Jeff Good, and Raktim Mukhopadhyay. Drag, drop, merge: A tool for streamlining integration of longitudinal survey instruments. HILDA, 2024.

[2] Dan Zhang, Yoshihiko Suhara, Jinfeng Li, Madelon Hulsebos, Çağatay Demiralp, and Wang-Chiew Tan. Sato: Contextual semantic type detection in tables. *PVLDB*, 13(12):1835–1848, 2020.

*Presenter