Making Table Understanding Work in Practice

Madelon Hulsebos* Sigma Computing

Sneha Gathani† Sigma Computing

James Gale Sigma Computing

Isil Dillig University of Texas

Paul Groth University of Amsterdam

Çağatay Demiralp Sigma Computing

Understanding the semantics of tables at scale is crucial for tasks such as data integration, preparation, and search. Table understanding methods aim at detecting a table's topic, semantic column types, column relations, or entities. With the rise of deep learning, researchers have developed models with excellent accuracy on benchmarks [1]. However, we observe a gap between model performance and their applicability in practice. Commercial data systems like Trifacta and Google Data Studio, for example, seem to primarily rely on simpler methods like regular expression matching for detecting a limited set of semantic types.We address the question: what do we need for table understanding models to work in practice?

Customization. Discussions with several medical companies highlighted the need to customize models for tables and semantics specific to their domains. Doing so requires new large datasets to represent these domains, costly labeling and long retraining procedures. Although the finetuning paradigm adopted for table understanding models aims to relieve these burdens [1], these models are not straightforward to finetune and still require large amounts of labeled data.

As demonstrated in information retrieval systems, an efficient way to accomplish model adaptation is to make models learn from interactive feedback. Such interactions should take minimal time and input, to maximize the effectiveness of the feedback cycle. To achieve this, our system, SIGMATYPER, builds on the data programming by demonstration framework [2]. We infer labeling functions from a table based on the user's feedback, implicit and explicit, and use these functions to generate new domain-specific training data.

Relevant tables. Another limitation of pretrained table understanding models for deployment in enterprises is that they are pretrained on tables that poorly resemble typical database tables as the training data mostly reflects tables found on the web [1, 4]. Therefore, tuning a pretrained model towards a representative data distribution and labels still takes too many resources that are often unavailable.

To overcome the data gap, new data sources [3, 5] aim at capturing database-like tables. We use GitTables [3] to train SIGMATYPER. For the subtask of semantic column type detection, we need semantic types common in the enterprise, science institutes, the medical domain, and beyond. Table columns in GitTables are annotated with over 1K semantic types from different ontologies. We select annotations from DBpedia to leverage knowledge base lookups of table entities.

Reliable inference. The quality and confidence of the inferences made with these models are typically unstable across labels, limiting the reliability of the model output. To date, no finetuning procedures are proposed to reflect whether samples are far from the training set. In practice, it is important to accurately reflect out-of-distribution data points to ensure high precision as errors are costly.

To overcome these challenges, systems are preferably hybrid: combining pragmatic, fast and transparent heuristics with learned models that offer high capacity, semantic coverage and out-of-distribution training techniques. SIGMATYPER implements a 3-step pipeline to infer semantic column types based on a table's header, column values, and embedding. Each step is executed only if a preset confidence threshold is not met by the prior step. SIGMATYPER yields the top-k semantic types for each column along with a confidence score to provide alternative suggestions in case of errors.

Future steps. We propose deeper analysis of the boundaries of pretrained models based on different customization approaches and real-world data to elicit future developments. Deficits should inform the next generation of table models, which likely build on table-specific representations. Beyond table representations, we believe that table semantics are also embodied by what people do with it which opens opportunities to incorporate actions into these models.

*Also with University of Amsterdam.

†Also with University of Maryland.

12th Annual Conference on Innovative Data Systems Research (CIDR'22), January 9-12, 2022, Chaminade CA, USA.

REFERENCES

[1] X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu. TURL: Table understanding through representation learning. *PVLDB*, 2020.

[2] S. Evensen, C. Ge, D. Choi, and Ç. Demiralp. Data programming by demonstration: A framework for interactively learning labeling functions. *arXiv preprint arXiv:2009.01444*, 2020.

[3] M. Hulsebos, Ç. Demiralp, and P. Groth. GitTables: A large-scale corpus of relational tables. *arXiv preprint arXiv:2106.07258*, 2021.

[4] S. Langenecker, C. Sturm, C. Schalles, and C. Binnig. Towards learned metadata extraction for data lakes. *BTW*, 2021.

[5] A. Vogelsgesang, M. Haubenschild, J. Finis, A. Kemper, V. Leis, T. Mühlbauer, T. Neumann, and M. Then. Get Real: How benchmarks fail to represent the real world. In *Proceedings of the Workshop on Testing Database Systems*, 2018.