Table Understanding in Practice

Madelon Hulsebos*
University of Amsterdam
m.hulsebos@uva.nl

We understand images, videos, and natural language, but tables are still a mystery. As a result, technological advances stay behind to intelligibly navigate, manage and exploit tables, while they pile up in databases. Lyft's data discovery system, Amundsen [3], and Google's Dataset Search [1], for example, still rely on manually provided table metadata. The practical success of deep learning stimulates the development of table understanding models to close this gap.

With models like Sherlock [6] and TURL [2], we can obtain table embeddings that encode their schema, relations, and entities. Evaluations of these models show impressive accuracy for generating metadata to facilitate tasks like data search and integration. These experiments, however, illustrate performance within an isolated context. We observed that actual deployment of table understanding models beyond this context is still challenging for various reasons.

First, existing datasets fail to represent offline tables. Existing table corpora, used to train table understanding models, primarily contain tables extracted from HTML pages, limiting the capability to represent offline database tables. To train and evaluate high-capacity models for applications beyond the web we need additional resources with tables that resemble relational database tables.

In this talk, we discuss GitTables [4], a corpus of currently 1.7M relational tables extracted from GitHub. Our continuing curation aims at growing the corpus to at least 10M tables. We annotated table columns in GitTables with more than 2K different semantic types from Schema.org and DBpedia. The corpus is available at https://gittables.github.io.

Analyses of GitTables showed that its structure, content, and topical coverage differ significantly from existing table corpora. We found that a semantic type detection model trained on GitTables obtains high prediction accuracy while the same model trained on tables from the web generalizes poorly. GitTables is currently being explored for developing systems for data discovery in data lakes, customizable table understanding models, and benchmarks for data integration.

Second, systems hardly adapt to unknown contexts. Training models on representative tables, is a necessary first step. But even if we train models on datasets with diverse coverage they may still encounter unknown semantics. Table understanding systems should, therefore, effectively adapt to the diverse contexts in which they are deployed.

Information retrieval systems have shown that iteratively learning from feedback is an efficient way to accomplish this. Such interactive feedback should be easy, taking minimal time and input, to maximize the effectiveness. Existing systems like Talend and Google Data Studio, however, often require complicated manual configurations for adaptation.

In this talk, we also present SigmaTyper [5], which builds on the Programming by Example framework [7] to implement domain adaptation. While it is trained on GitTables, it enables implicit and explicit feedback on annotations to adapt to unknown semantics. This feedback is, in turn, used to infer labeling functions. These functions can be exploited in different ways to customize the system for different contexts. In SigmaTyper, they are used to generate new training data using Data Programming.

We are not there, yet. Table understanding models facilitate many table tasks like data search and integration. In order to effectively deploy these models in practice, we need to solve some remaining challenges first. To start, we need to identify the blind spots of concurrent table understanding models by gaining a deeper understanding of them. Deficits should inform the next generation of table models. We envision that these models will rely on table-specific encodings instead of representations of natural language.

REFERENCES

[1] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *The World Wide Web Conference*. 1365–1375.

[2] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2020. TURL: Table understanding through representation learning. *arXiv preprint arXiv:2006.14806* (2020).

[3] Mark Grover. 2019. Amundsen — Lyft's data discovery & metadata engine. https://eng.lyft.com/amundsen-lyfts-data-discovery-metadata-engine-62d27254fbb9

[4] Madelon Hulsebos, Çağatay Demiralp, and Paul Groth. 2021. GitTables: A Large-Scale Corpus of Relational Tables. *arXiv preprint arXiv:2106.07258* (2021).

[5] Madelon Hulsebos, Sneha Gathani, James Gale, Isil Dillig, Paul Groth, and Çağatay Demiralp. 2021. Making Table Understanding Work in Practice. *arXiv preprint arXiv:2109.05173* (2021).

[6] Madelon Hulsebos, Kevin Hu, Michiel Bakker, Emanuel Zgraggen, Arvind Satyanarayan, Tim Kraska, Çagatay Demiralp, and César Hidalgo. 2019. Sherlock: A deep learning approach to semantic data type detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1500–1508.

[7] Henry Lieberman. 2001. *Your wish is my command: Programming by example*. Morgan Kaufmann.

*Also with Sigma Computing.