Structural Embedding of Data Files with MAGRITTE

Gerardo Vitagliano
Hasso Plattner Institute
University of Potsdam, Germany
gerardo.vitagliano@hpi.de

Mazhar Hameed
Hasso Plattner Institute
University of Potsdam, Germany
mazhar.hameed@hpi.de

Felix Naumann
Hasso Plattner Institute
University of Potsdam, Germany
felix.naumann@hpi.de

Abstract

Large amounts of tabular data are encoded in plain-text files, e.g., CSV, TSV and TXT. Plain-text formats allow freedom of expression and encoding, fostering the use of non-standard syntaxes and dialects. Before analyzing the content of such files, it is necessary to understand their *structure*, e.g., recognize their dialect [8], extract metadata [6, 4], or detect tables [1]. Previous work on table representation focused on learning the *semantics* of data cells [5, 2, 9], with the assumption that the syntactical properties of a file are known to end users.

We propose MAGRiTTE, an approach to synthetically represent the structural features of a data file. MAGRiTTE is a self-supervised machine learning model trained to learn structural embeddings from data files. The architecture of MAGRiTTE is composed of two components. The first is a transformer-encoder architecture, based on BERT [3] and pre-trained to learn row embeddings. The second is a DCGAN-autoencoder [7] trained to produce file-level embeddings. To pre-train the transformer architecture on structural features, we propose two core adaptations: a novel tokenization stage and specialized training objectives. To abstract the data content of a file, and train the transformer architecture on structural features, we introduce "pattern tokenization": assuming that structural properties are identifiable through special characters, we reduce all alphanumeric characters to a set of few general patterns. After tokenization, the rows of the input files are split on newline characters and a percentage of the special character tokens is masked before feeding it to the row encoder model. The row-transformer model is then trained on two objectives, reconstructing the masked tokens, and identifying whether pairs of rows belong to the same file. The row embeddings produced by this model are then used as the input for the file embedding stage of MAGRiTTE. In this stage, the generator and discriminator models are trained on the row embeddings feature maps. As the file-wise embedding vector, we use the output features produced from the last layer of the encoder.

We shall evaluate the effectiveness of our learned structural representations on three tasks to analyze unseen data files: (1) fine-grained dialect detection, i.e., identifying the structural role of characters within rows, (2) line and cell classification, i.e., identifying metadata, comments, and data within a file, (3) table extraction, i.e., identifying the boundaries of tabular regions. We compare the use of MAGRiTTE encodings with state-of-the-art approaches that were specifically designed for these tasks. In future work, we aim at using MAGRiTTE embeddings to automatically perform structural data preparation, e.g., removing unwanted rows, or changing file dialects, as well as to index the content and structure of data files within data lakes.

Table Representation Learning Workshop at NeurIPS 2022

Acknowledgments and Disclosure of Funding

This research was funded by the HPI Research School on Data Science and Engineering.

References

[1] C. Christodoulakis, E. Munson, M. Gabel, A. D. Brown, and R. J. Miller. Pytheas: Pattern-based table discovery in CSV files. *PVLDB*, 13(11):2075–2089, 2020.

[2] X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu. TURL: table understanding through representation learning. *PVLDB*, 14(3):307–319, 2020.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, volume 1, pages 4171–4186. Association for Computational Linguistics, 2019.

[4] M. Hameed, G. Vitagliano, L. Jiang, and F. Naumann. SURAGH: syntactic pattern matching to identify ill-formed records. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*, pages 2:143–2:154. OpenProceedings.org, 2022.

[5] H. Iida, D. Thai, V. Manjunatha, and M. Iyyer. TABBIE: Pretrained Representations of Tabular Data. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3446–3456. Association for Computational Linguistics, 2021.

[6] L. Jiang, G. Vitagliano, and F. Naumann. Structure detection in verbose CSV files. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*, pages 193–204. OpenProceedings.org, 2021.

[7] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Y. Bengio and Y. LeCun, editors, *International Conference on Learning Representations, ICLR*, 2016.

[8] G. J. J. van den Burg, A. Nazábal, and C. Sutton. Wrangling messy CSV files by detecting row and type patterns. *Data Mining and Knowledge Discovery*, 33(6):1799–1820, 2019.

[9] P. Yin, G. Neubig, W. Yih, and S. Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. In *ACL*, pages 8413–8426. Association for Computational Linguistics, 2020.