A Framework for Automatically Interpreting Tabular Data at Orange

Yoan Chabot, Pierre Monnin, Pierre Monnin, Frédéric Deuzé, Viet-Phi Huynh, Thomas Labbé, Jixiong Liu, Raphaël Troncy

A Framework for Automatically Interpreting Tabular Data at Orange

Yoan Chabot[1], Pierre Monnin[1], Frédéric Deuzé[1], Viet-Phi Huynh[1], Thomas Labbé[1], Jixiong Liu[1,2], and Raphaël Troncy[2]

[1] Orange, France
yoan.chabot@orange.com
[2] EURECOM, France
raphael.troncy@eurecom.fr

On the Importance of Interpreting Tabular Data

Large parts of knowledge of companies are encoded in tabular data. Being able to interpret such data is key to increase business efficiency and to propose innovative services, and Orange is no exception. With more than 140,000 employees worldwide and a heterogeneous client portfolio, Orange produces a phenomenal amount of tabular data every day. These tables are viscerally embedded in internal services and products (*e.g.*, network logs, multimedia catalogs). Hence, they are a source to discover new knowledge. Although this encourages the development of efficient tools to process them, several issues are negatively impacting their use. First, the volume curse makes difficult to identify the right dataset for a given use case. Then, the knowledge gap between data producers/consumers is exacerbated by our language footprint (seven main languages), the heterogeneous tools producing various table formats, and the experience/jobs of employees leading to similar concepts being expressed by different terms across tables.

Making Sense of Tables: DAGOBAH

One lever to address these challenges resides in Semantic Table Interpretation (STI), *i.e.*, making tabular data intelligently processable by matching elements of tables and constituents of Knowledge Graphs (*e.g.*, Wikidata or specific domain/enterprise ones). Providing semantic annotations is a key asset for search/recommendation engines and natural language based services. Moreover, KGs can be used to drive STI while being themselves enriched by STI's results. Indeed, transferring knowledge to KGs allows previously dormant knowledge to be structured and queried efficiently. In this view, Orange/EURECOM research teams have developed DAGOBAH, a framework interpreting tables automatically. DAGOBAH is packaged in a RESTful API named TableAnnotation on

the Orange Developer portal[3]. DAGOBAH offers four features: *i)* preprocessing (data cleaning, primitive typing, table orientation, etc.), *ii)* Cell-Entity Annotation associating each cell with a KG entity, *iii)* Column-Type Annotation typing each column with a KG concept, and *iv)* Columns-Property Annotation identifying properties between table columns. It comes also in two flavours: DAGOBAH SL based on a smart lookup of entities which are then disambiguated by comparing the context given by the table row under study with the context of the candidates in the KG; DAGOBAH Embeddings exploiting geometric properties of vector spaces to cluster candidates. While computationally more expensive, embeddings provide promising results on highly ambiguous tables. These two approaches have proven to be competitive during the SemTab challenge [1, 3].

Future Work

Interpreting tables has become a crucial task attracting a lot of attention in recent years and the SemTab challenge [4] provides a relevant forum for comparing approaches. However, there is still a long way to go to support the richness of tables stored in company repositories. In particular, two limitations have been identified in the current gold standards: 1) tables are generally homogeneously shaped and the noise artificially generated does not represent the real world complexities spectrum, and 2) target KGs are mostly encyclopaedic or limited to very specific domains. A challenge will be to propose datasets encoding more difficulties such as matrix tables or multi-values cells. Another challenge Orange will face is the application of annotation techniques to tables derived from business knowledge. This requires the development of Orange's own KG that focuses on its activities, following the accelerating trend of Enterprise KGs. This raises difficult questions such as how to support our various business areas and how to bootstrap such a KG from publicly available KGs? It also requires a complex pipeline as demonstrated by Amazon's Product Graph [2]. To this aim, Orange's future research work will be structured along the harvesting of the company's tabular data, the complementarity of text/tables in the knowledge extraction process, and the reconciliation of knowledge in a high quality KG.

References

1. Chabot, Y., et al.: DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System. In: SemTab 2019. pp. 41–48 (2019)
2. Dong, X.L., et al.: AutoKnow: Self-driving knowledge collection for products of thousands of types. In: 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2724–2734 (2020)
3. Huynh, V.P., et al.: DAGOBAH: Enhanced Scoring Algorithms for Scalable Annotations of Tabular Data. In: SemTab 2020. pp. 27–39 (2020)
4. Jiménez-Ruiz, E., et al.: Results of semtab 2020. In: CEUR Workshop Proceedings. vol. 2775, pp. 1–8 (2020)

[3] https://developer.orange.com, API available upon invitation.