# Tab2Onto: Unsupervised Semantification with Knowledge Graph Embeddings

Hamada M. Zahera[1], Stefan Heindorf[1], Stefan Balke[3], Jonas Haupt[2], Martin Voigt[2], Carolin Walter[4], Fabian Witter[3], and Axel-Cyrille Ngonga Ngomo[1]

[1] DICE Group, Paderborn University, Germany
{hamada.zahera, heindorf, axel.ngonga}@upb.de
[2] elevait GmbH & Co. KG, Dresden, Germany
{jonas.haupt, martin.voigt}@elevait.de
[3] pmOne AG, Paderborn, Germany
{stefan.balke, fabian.witter}@pmone.com
[4] USU Software AG, Karlsruhe, Germany
carolin.walter@usu.com

**Abstract.** A large amount of data is generated every day by different systems and applications. In many cases, this data comes in a tabular format that lacks semantic representation and poses new challenges in data modelling. For semantic applications, it then becomes necessary to lift the data to a richer representation, such as a knowledge graph that adheres to a semantic ontology. We propose Tab2Onto, an unsupervised approach for learning ontologies from tabular data using knowledge graph embeddings, clustering, and a human in the loop. We conduct a set of experiments to investigate our approach on a benchmarking dataset from a medical domain and learn the ontology of diseases. Our code and datasets are provided at https://tab2onto.dice-research.org/

**Keywords:** Ontology Learning · Tabular Data · Knowledge Graph Embeddings · Human-In-The-Loop

## 1 Introduction

Data-driven companies collect large amounts of data from various sources to improve their business analytic and decision-making processes. In most cases, this data comes in a tabular format (e.g., as CSV files). The lack of semantic information in tabular data leads to machines often being unable to assign unique semantics to their content.

*Semantification* [2] is the process of converting data into a representation with unique semantics, e.g., an RDF knowledge graph, that tackles the aforementioned drawback of tabular data. It also simplifies data integration [5] and explainable machine learning [3]. However, current semantification frameworks rely on numerous hand-crafted scripts, which require expensive maintenance by IT service providers. We propose the Tab2Onto approach, an unsupervised semantification process which exploits knowledge graph (KG) embeddings. Our

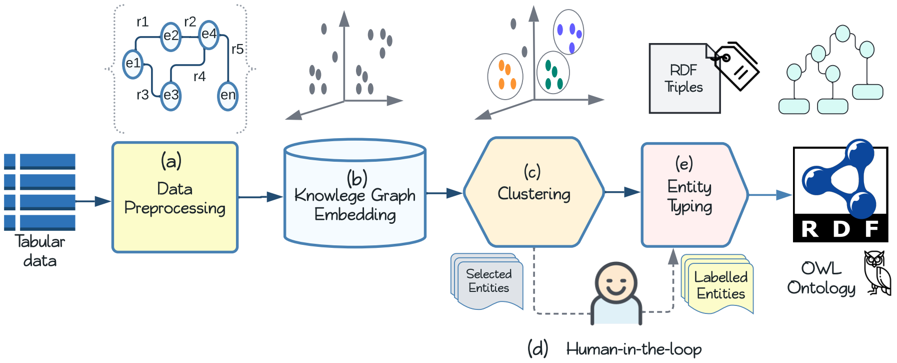

**Fig. 1:** Tab2Onto pipeline for semantification.

approach works as follows (see Figure 1): (i) construct a KG from tabular data, (ii) employ KG embeddings to represent entities and relations, (iii) apply hierarchical, unsupervised clustering, (iv) have a human in the loop to assign labels for the computed clusters, and (v) generate an ontology.

## 2 Related Work

Recently, many approaches have been proposed to construct ontologies from *textual* data. We refer to the survey paper [8] for more details about ontology learning from text. Few studies on constructing ontologies from *tabular* data (e.g., CSV, spreadsheets) have been carried out in recent research. For example, the authors of [2] propose a *user-driven* approach that requires considerable manual work. The approach in [4] only populates an existing ontology from tabular data. Furthermore, the work presented in [6] demonstrates the significance of transforming tabular data into RDF to capture semantic information. The authors propose an ontology-driven approach for generating RDF from multiple CSV files. However, they assume that each CSV file contains entities from the same domain, which is not the case for most real-world data. To deal with entities from different domains, we use entity clustering to group similar entities together. To the best of our knowledge, this is the first attempt that combines KG embeddings, clustering, and a human in the loop for ontology learning.

## 3 Approach

Our approach takes a single[5] CSV file as input and generates an OWL ontology as output. Figure 1 shows the pipeline of the Tab2Onto approach, including five steps. In the *data preprocessing* step (Fig. 1 a), we convert the input data to an RDF graph using the Vectograph library[6] that transforms each cell entry $e_{i,j}$

[5] In case of multiple CSV files, they are joined into a single file.

[6] https://github.com/dice-group/Vectograph

**Table 1:** Clustering for type prediction on FB15k-237. Best results in bold.

| **Algorithm** | **TransE** | | **DistMult** | | **RotatE** | | **QMult** | |
|---|---|---|---|---|---|---|---|---|
| | **Acc.** | **$F_1$** | **Acc.** | **$F_1$** | **Acc.** | **$F_1$** | **Acc.** | **$F_1$** |
| K-Means | 0.784 | 0.751 | 0.771 | 0.741 | 0.282 | 0.200 | **0.785** | **0.803** |
| Agglomerative | 0.779 | 0.746 | 0.781 | 0.749 | 0.284 | 0.201 | 0.744 | 0.775 |
| HDBSCAN | 0.678 | 0.624 | 0.475 | 0.362 | 0.276 | 0.119 | 0.276 | 0.119 |

in row $i$ and column $j$ to a <*subject-predicate-object*> triple, i.e., ($e_i$, $e_j$, $e_{i,j}$) where $e_i$ denotes the name of row $i$ and $e_j$ the name of column $j$. Further, we represent entities and relations in the RDF graph using KG embeddings (Fig. 1 b). Each entity and relation is represented as a $d$-dimensional vector ($\mathbb{R}^d$) in the embedding space, where similar entities are close to each other. In the clustering step (Fig. 1 c), we use the K-Means algorithm to identify clusters of entities. Each cluster contains a set of entities with similar properties and common relations. In the next step (Fig. 1 d), our goal is to assign labels (i.e., classes or types) to the clustered entities. For this purpose, we employ a *human in the loop* to assign one label to each cluster. We ask said human to specify labels for a few entities from each cluster. For each cluster, we sample some entities close to its centroid and present a set of RDF triples about these entities via a web interface. The user can then manually assign a label to each entity. After that, we *propagate* the major label to all entities within the same cluster (Fig. 1 e). Finally, we construct an OWL ontology based on the labelled clusters using the OWLready2 library.[7] The learned ontology contains a taxonomy of OWL classes and entities (i.e., OWL individuals) with type information.

## 4 Experiments

We aim to answer the following questions: **(Q1)** *Which KG embeddings yield the best clusters of entities in the embedding space?* **(Q2)** *Which clustering approach yields the best clusters of entities?* **(Q3)** *How well does our pipeline work for the semantification of tabular data?*

**Evaluation Setup:** For research questions **Q1** and **Q2**, we use the popular KG benchmark *FB15k-237* with types such as `movie`, `person`, etc. The dataset includes a subset of the Freebase Knowledge Graph with $14,951$ entities and 237 relations. For **Q3**, we use the *Lymphography*[8] dataset, which contains tabular data about 148 instances of lymphography diagnoses with 18 attributes. As metrics, we use *accuracy* and *macro*-$F_1$ to evaluate the predicted types of entities compared to the ground-truth types in the *FB15k-237* dataset. Furthermore, we use Evolearner [3] from the Ontolearn library to evaluate the generated ontology.

[7] https://github.com/pwin/owlready2

[8] https://archive.ics.uci.edu/ml/datasets/Lymphography

**Table 2:** Tab2Onto semantification of *Lymphography* with QMult embeddings and K-Means clustering.

| **Approach** | **Acc.** | $\mathbf{F_1}$ |
|---|---|---|
| Tab2Onto (*unsupervised*) | **0.666** | **0.728** |
| Random (*unsupervised*) | 0.533 | 0.485 |
| Logistic regression (*supervised*) | 0.833 | 0.818 |

### 4.1 Embedding-based Clustering for Type Prediction

To answer **Q1**, we experimented with the KG embeddings TransE, DistMult, RotatE, and QMult [1]. Table 1 shows the evaluation results in terms of *accuracy* (**Acc.**) and *macro*-$F_1$ ($\mathbf{F_1}$) measure. Our results demonstrate that QMult embeddings achieve superior performance over other embedding models, with an $F_1$-score of 0.803 compared to 0.751 by TransE (K-Means clustering).

To answer **Q2**, we evaluated the performance of different clustering methods: K-Means, agglomerative clustering, and HDBSCAN. Table 1 reports our evaluation results for each method with the KG embedding models used in **Q1**. We observe that K-Means achieves the best performance in clustering entities. In particular, K-Means outperforms agglomerative clustering by absolute +0.028 in terms of $F_1$-score (for QMult embeddings). Based on these findings, we employ the best combination of KG embeddings (QMult) and clustering algorithm (K-Means) in the full pipeline of our approach in the next section.

### 4.2 Semantification of Tabular Data

To answer **Q3**, we investigated the application of our pipeline in the medical domain. We used the benchmarking dataset *Lymphography*, which provides lymphograms and their attributes as tabular data (e.g., *lymphatics, lymNodesEnlar, defectInNode, extravasates*). Our goal is to infer types of lymphatic diseases (*Normal, Fibrosis, Metastases, Malign-Lymph*) and represent them as OWL classes in the generated ontology. Starting from tabular data, we apply the full pipeline of Tab2Onto as follows: we transform the tabular data of Lymphography into an RDF graph in step (a); then we learn QMult embeddings in step (b); we cluster entities using the K-Means approach in step (c); we employ a human in the loop to assign labels (*Normal, Fibrosis, Metastases or Malign-Lymph*) to a set of sampled entities from each cluster in step (d). Finally, the output of Tab2Onto is an ontology that contains a taxonomy of OWL classes based on the cluster labels, in step (e).

To evaluate the predicted lymphatic types, we compared our *unsupervised* approach to *random-labelling* with probabilities reflecting the class distribution. Further, we used *supervised* logistic regression as an upper-bound baseline for type prediction. Table 2 shows that Tab2Onto outperforms *random-labelling* with a large margin, up to +0.13 accuracy and +0.24 *macro*-$F_1$ scores; we are reasonably close to the *supervised* logistic regression approach. In addition, we evaluated the application of the generated ontology in a concept learning task. Using the positive and negative examples of the Lymphography dataset in SML-Bench [7], the state-of-the-art concept learner EvoLearner [3] learns a concept with an $F_1$-score of 0.82 on the automatically generated ontology compared to 0.84 for a concept learned on SML-Bench's manually created ontology.

## 5 Conclusion

We present Tab2Onto, an unsupervised semantification approach for learning an ontology from tabular data without requiring any labelled training data. Our approach clusters entities using their KG embeddings to derive their types. By employing embedding-based clustering and a human in the loop, our approach can efficiently convert tabular data into a machine-readable format that can be linked to knowledge graphs. In future work, we will explore density-based clustering with further hyperparameter tuning. We will also conduct more experiments with semi-supervised approaches to learn ontology with few labelled data.

**Acknowledgements** This work has been supported by the German Federal Ministry of Education and Research (BMBF) within the project DAIKIRI under grant number 01IS19085B and the German Federal Ministry for Economic Affairs and Climate Action (BMWK) within the project RAKI under grant number 01MD19012B.

## Bibliography

[1] Demir, C., Moussallem, D., Heindorf, S., Ngomo, A.C.N.: Convolutional hypercomplex embeddings for link prediction. In: Asian Conference on Machine Learning, pp. 656–671, PMLR (2021)

[2] Ermilov, I., Auer, S., Stadler, C.: User-driven semantic mapping of tabular data. In: I-SEMANTICS, pp. 105–112, ACM (2013)

[3] Heindorf, S., Blübaum, L., Düsterhus, N., Werner, T., Golani, V.N., Demir, C., Ngomo, A.C.N.: Evolearner: Learning description logics with evolutionary algorithms. In: WWW (2022)

[4] Nederstigt, L.J., Aanen, S.S., Vandic, D., Frasincar, F.: FLOPPIES: A framework for large-scale ontology population of product information from tabular data in e-commerce stores. Decis. Support Syst. **59**, 296–311 (2014)

[5] Ngonga Ngomo, A.C., Sherif, M.A., Georgala, K., Hassan, M.M., Dreßler, K., Lyko, K., Obraczka, D., Soru, T.: Limes: A framework for link discovery on the semantic web. KI-Künstliche Intelligenz **35**(3), 413–423 (2021)

[6] Sharma, K., Marjit, U., Biswas, U.: Automatically converting tabular data to rdf: An ontological approach. Int J Web Semant Technol (2015)

[7] Westphal, P., Bühmann, L., Bin, S., Jabeen, H., Lehmann, J.: Sml-bench - A benchmarking framework for structured machine learning. Semantic Web **10**(2), 231–245 (2019)

[8] Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text: A look back and into the future. ACM Comput. Surv. **44**(4), 20:1–20:36 (2012)