Towards Transforming Tabular Datasets into Knowledge Graphs

Nora Abdelmageed[0000−0002−1405−6860]
(Early Stage PhD)

Heinz Nixdorf Chair for Distributed Information Systems
and Computer Vision Group
Michael Stifel Center Jena, Germany
Friedrich Schiller University Jena, Germany
nora.abdelmageed@uni-jena.de

Abstract. Many applications rely on the existence of reusable data. The FAIR principles identify rich descriptions of data and metadata as the key ingredients for achieving reusability. However, creating descriptive data requires massive manual effort. One way to ensure that data is reusable is by integrating it into a Knowledge Graph (KG). The semantic foundation of these graphs provides the necessary description for reuse. In this paper, we focus on tabular data and how that can be integrated into a KG. Besides the tabular data itself, we leverage existing metadata and publications describing the datasets for the KG construction. To tackle this task, we introduce a machine-learning based framework. Our framework consists of three core modules. The first module predicts the concepts of the KG from various data sources. In the second module, we extract possible relations among these concepts. Afterwards, we will integrate the two modules to build the final KG. As an example domain to develop and evaluate our approach, we focus on Biodiversity research. This is a data-rich domain with a particularly high need for data reuse. We present preliminary results in the context of building a KG schema given table headers. We cluster these headers using two types of representations, word embeddings, and syntactic representation. Our results show that embeddings can catch high-level semantics of headers; thus, they are better descriptors.

Keywords: Knowledge Graph Construction · Table Understanding · Named Entity Recognition (NER) · Relation Extraction (RE).

1 Introduction

Recently, Knowledge Graphs (KGs) have become popular as a means to represent a domain knowledge. Auer et al. [2] propose them as a way to bring scholarly communication to the 21st century. In the Open Research KG, they propose to model artifacts of scientific endeavors, including publications and their key messages. Datasets supporting these publications are important carriers of scientific knowledge and should thus be included in KGs. An important side effect of this inclusion is that it supports FAIRness of data [23]. The FAIR principles identify rich descriptions as the major prerequisite for reusabilty. Since KGs make the semantics of the data explicit, they provide these rich descriptions. It is not trivial, however, to add datasets to KGs by manual transformation is prohibitively expensive. In our work, we aim to enable (semi-)automatic

integration of information from tabular datasets into KGs. We will exploit the datasets themselves, but also auxiliary information like existing metadata and the associated publications. To address this problem, we will combine semantic web technologies and machine learning techniques. In this way, we can extend and enrich existing KGs. We will develop and test the proposed approach using datasets from Biodiversity research. This is an area of science of particular societal importance and a field with a strong need for data reuse, e.g., by KGs.

As a basis for our work, we held several meetings with Biodiversity scientists. We found that Biodiversity synthesis work is done today as follows: the research team searches for all datasets relevant to their research question. This happens via searches in data repositories, literature search and personal connections. The publications found are then read to find essential references. Metadata about datasets is extracted from them and, in the case of data repositories, the information uploaded. All of this information is then manually collated. This serves as a basis to decide on which data is usable for the study at hand, which conversions and error corrections are necessary and how the data can be integrated. This process can take several months. Providing well described data in a KG would drastically reduce the required effort.

Various solutions aim at domain-specific KG construction exist. Page [18] shows guidelines for the construction of a Biodiversity KG. However, the resultant KG is coarse-grained. For example, the author proposes linking a whole dataset to a publication and an author. A more fine-grained solution, a rule-based framework [4] constructs a Biodiversity KG from publication text. It covers both named entity recognition and relation extraction tasks. The authors use different types of taggers to capture a wide range of information inside the document. A similar approach is taken in [14] with a broader goal of information extraction from textual scientific data in general. At this point, there is a broad interest in building scientific KGs evidenced by the existing approaches. However, none of them deal efficiently with tabular datasets yet.

The rest of the paper is organized as follows: Section 2 covers the main categories of the related work. The problem statement, research questions and contributions are mentioned in Section 3. Section 4 discusses the research strategy. Section 5 presents the evaluation strategies. Preliminary experiments and results are outlined in Section 6. Finally, we conclude and discuss future work in Section 7.

2 State of the Art

In this section, we cover the essential related work needed for our proposed framework in the following subsections: i) Understanding tabular data, and ii) Understanding textual data. Both focus on obtaining entities of interest and their relations from tables and publication text respectively.

2.1 Understanding Tabular Data

Two classes of approaches address the tabular dataset understanding. The first category aims at matching the values of table cells, columns and column-column to KG entities, classes and properties. The second category, learning semantic table properties, allows to predict a general column's class that might not exist in knowledge base (KB).

Table Cells and Columns to KG Matching In this category, the current works aim at creating ontology mappings for the table on various levels, like cells, columns, and column-column relations. For example, a cell has a value *Germany* will be linked to *dbr:Germany*, while a column contains a list of countries will be mapped to *dbo:Country*. Both annotations are DBpedia [1] entity and class. There are two possibilities to achieve this task. First, there are semi-automatic approaches, which involve

human intervention. Karma [10], for instance, provides recommendations for ontology mappings or lets users define a new mapping. These methods are very time-consuming. Second, there are fully-automatic techniques, which do not require any manual effort. The Sem-Tab challenge[1], which took place for the first time at ISWC 2019, presents three different tasks for the automatic approaches: i) Cell Entity Assignment (CEA) matches a cell of the table to a KG entity. ii) Column Type Assignment (CTA) assigns a KB class to a column. iii) Column Property Assignment (CPA) selects relations between two different columns. All the presented works use the solution of the CEA as a core part of solving the others. So for our discussion, we will focus on the CEA task.

MTab [16] relies on the brute force lookup of all table signals, then applies majority voting as a selection criterion. This technique achieves the best results, but it is computationally expensive and does not suite real-world systems. Another approach introduced is Tabularisi [21]. It also looks up the KB services, but it converts the results into a feature space using TF-IDF[2]. Then, the final decision is the top-1 value. Finally, DAGOBAH [5] searches the entities in the vector space model of the KB. Then, it applies the K-means clustering algorithm on the embeddings. Finally, it selects the cluster with the highest score to assign the entity type. However, the performance of these works presented will decrease when there are missing or inaccurate mappings in the KBs. We define such a problem as a knowledge gap. In other words, a problem appears when the dataset and the target KB are not derived from the same distribution, from e.g. DBpedia. In our scope, which targets Biodiversity datasets, we need a way to infer types and discover new entities and relationships that co-occurred in real data and might be missing in the KB.

Table Semantic Properties Learning These approaches capture the semantic structure of the table. They heavily depend on machine learning. An exciting work, TabNet [17], classifies the Genuine web tables[3] into one of several predefined categories. TabNet relies on a hybrid neural network to learn both the inter-cell and high-level semantics of the whole table. We consider using TabNet in our work as a preprocessing step, such that, it can filter the input tables. A further exciting work [6] introduces ColNet for learning the semantic type of a table column. In the prediction part, they combine the pure prediction by the network with the majority voting by the lookup services. They achieve the best results using an ensemble strategy. By this means, ColNet proposes a solution for the knowledge gap which exists in the DBpedia KB. In our context, we plan to start from this architecture for column type prediction and extend it to our domain.

2.2 Understanding Textual Data

As we will combine data from publications, we discuss both Named Entity Recognition (NER) and Relation Extraction (RE) as the most important techniques for natural language processing. According to [12], NER involves the extraction of mentioned entities in natural language text and their classification in predefined categories. The authors also present a framework of a typical NER system. They divide the approaches to NER into two main categories. 1) Traditional approaches, including rule-based, unsupervised, and feature-based supervised approaches. For all of these techniques, selecting meaningful features remains a crucial problem. 2) Deep learning techniques can solve

[1] `http://www.cs.ox.ac.uk/isg/challenges/sem-tab/`

[2] Term Frequency-Inverse Document Frequency, a well known information retrieval metric, capturing importance of a term for a document.

[3] Tables that contain semantic triples, i.e., subject-predicate-object.

the mentioned problem by automatically selecting features by using, for example, a Recurrent Neural Network (RNN). An interesting domain-specific NER is Bio-NER [24]. Mainly, it leverages the input representation using word vectors by automatically learning them from unlabeled biomedical text. In this way, it solves the problem of feature engineering. Another exciting work leverages semantics from external resources [9]. The authors explore Wikipedia[4] as an external source of knowledge to improve the performance of NER. They extract the first part of the corresponding Wikipedia entry and the category labels from it. For better input representation, these category labels are added to the engineered features.

Relation Extraction techniques aim at semantic relations extraction between entity mentions and classification of them inside natural language text [3]. This classical survey covers both supervised and semi-supervised techniques for this task. However, it shows many limitations, like the overhead of the manual annotations and the criteria for selecting a good training seed. Moreover, such approaches are hard to extend and require new training data to detect new relations. However, [20] solves these problems by introducing the distant supervision technique. The authors also address the relation extraction as a supervised learning method but, without paying the cost of labeling the dataset. They leverage the KB as an external source of the existing relations. This technique has various challenges. It helps extend an existing KB but it is not useful to construct one from scratch.

3 Problem Statement and Contribution

We discuss in this section our main research problem and questions we aim to address, and our contributions.

3.1 Problem Statement & Research Questions

Our core research problem is how to enable the reusability of tabular data by adopting and extending machine learning techniques. For successful data reuse, a good, ideal machine-readable description of the data is essential. The desired descriptions are represented as a KG. However, today, creating such descriptions requires considerable manual effort. We believe that building such a KG automatically will only be possible by leveraging auxiliary information besides the dataset itself. The crucial sources of additional information are in our case metadata and publications. Our evaluations comparing the addition of these various data sources will show the correctness of this assumption.

Our research focuses on how to automate transformation of tabular datasets into a KG using various machine learning techniques. Since there is a massive amount of Biodiversity data available, we choose it as our first domain of interest. We divide this general research problem into three fine-grained research questions:

- **RQ1** - How can we use tabular datasets for KG construction?
- **RQ2** - How can we leverage the existing metadata in understanding the original dataset?
- **RQ3** - How can we benefit from the information in the associated publications to enrich the constructed KG?

3.2 Contributions

Our overall contribution will be to enable the automatic integration of tabular datasets into KGs thereby considerably increasing FAIRness, in particular reusability. This aim will be reached by several contributions:

[4] `https://www.wikipedia.org/`

- We will develop methods that take a tabular dataset as input and automatically create a KG out of it. These methods will determine the meaning of individual columns and their data type as well as relationships across columns. Such tools are useful to increase tabular data understanding even without the subsequent transformation in a KG.
- We will extend these tools to leverage potentially available auxiliary information, in particular metadata and publications describing the dataset.
- We will implement these methods into a framework.
- We will evaluate the individual methods as well as the overall system.

4 Research Methodology and Approach

In this section, we discuss research methodology pipeline and our proposed framework's conceptual model overview.

4.1 Research Methodology

Figure 1 shows our research methodology. In the first step of our pipeline, we conducted several meetings with domain experts from the Biodiversity field for requirement gathering as described in Section 1. Based on their requirements, we came up with three main stages for our project: Firstly, we will aim to build a KG from the tabular dataset itself as a standalone data source. Secondly, we will add information gained from metadata or any auxiliary semi-structured data. Finally, we do further extensions to the resultant KG using the related publications, either by the use of abstracts or the full texts. For each of these stages, we will perform a complete development cycle from an analysis of the state of the art and concept development to implementation, evaluation and publication. At this phase in the project, the evaluation will focus on performance metrics (see below). Once the complete system has been implemented, a final overall evaluation including a user study will be undertaken.

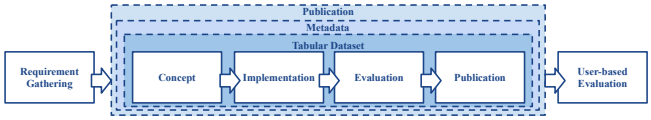


Fig. 1: Research methodology pipeline

4.2 Conceptual Model View

Figure 2 illustrates the architecture of our proposed framework in the first stage. It receives a tabular dataset in an Excel sheet or a CSV file as input. The framework transforms this tabular dataset into a full KG, such that the resultant KG has schema and instances inferred from the tabular dataset. In the country-city example, object entities [5], e.g., "Country", contribute graph nodes. However, non-object entities, e.g., "Area", contribute the relations. Our framework consists of three core modules: i) Concept Prediction; it predicts the KG schema class of a given column in the table. It also encapsulates various approaches like NER, lookup services, taggers, and classifiers (i.e., neural networks). ii) Relations Detection, on the one hand, finds a possible relation between two object columns by using a relation extraction technique (i.e., distant supervision). On the other hand, it looks up the domain knowledge for a relation between a concept and a non-object entity (e.g., "Country", "Area"). We will filter both concepts and relations candidates on specific criteria. iii) KG Construction builds the final full graph given the filtered concepts and relations with the original dataset.

[5] Object entities: are entities that can be a page on Wikipedia.

In Stage 2 and 3 of the project, we will leverage information about the dataset in other resources that are not existing in the tabular dataset itself. For example, a metadata file or a publication could have the unit of "Area" in km^2. By this means, we enrich the constructed KG from tabular data by the secondary information that exists in the other sources. This extension will require adaptations to all three core modules.

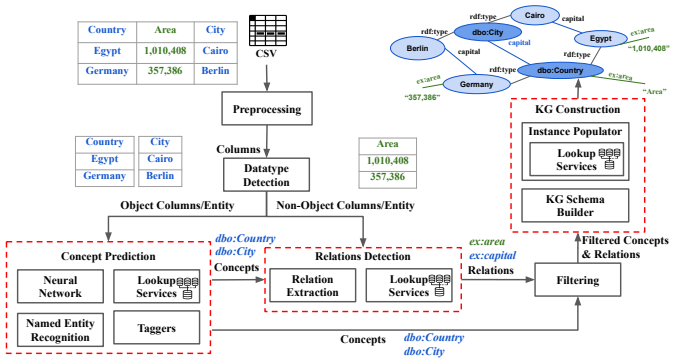


Fig. 2: System architecture of the proposed framework with a simple Country-City example in tabular dataset.

5 Evaluation Plan

We will need two types of evaluation for this work: Firstly, we will evaluate the performance of the framework and the effect of using additional information in Stages 2 and 3 using standard evaluation metrics. Secondly, we will evaluate the quality of the resulting KG with a user-based evaluation.

5.1 Performance Evaluation of the Framework

At the end of each stage, we will evaluate the performance of the framework and of individual modules using evaluation metrics like the standard Precision (Pr), Recall (R), and F-score. Besides these metrics, we can adopt others like Macro and Micro versions of them, especially when we have an unbalanced dataset or when implementing the natural language processing modules. These modules are multi-class classification tasks. Thus, we are interested in measuring the robustness of the system per each class. At first, this strategy gives an impression of the system performance after each step. Secondly, it enables agile development by dividing and focusing on each separate module as a standalone project. Additionally, for testing the first phase, we will use benchmark datasets: T2Dv2 [11], Limayae [13] and SemTab2019 Data Sets [8].

5.2 User-based Evaluation

We claim that transforming tabular datasets into a KG enhances their reusability. A user study is needed to examine whether that is indeed true. This user study will be performed at the end of Stage 3. We consider two possible options for this evaluation: First we will conduct an end-to-end assessment. It differs from the previous evaluation strategy. This assessment concerns the information encapsulated inside the KG itself. We can achieve this kind of evaluation by preparing a list of predefined questions and issuing them in the form of queries against the SPARQL endpoint of the constructed KG. Thus, the retrieved answers can be used as a metric. We will held this type of evaluation on the final KG using the tabular dataset and after including the metadata and finally after including the related publications. A second option would be to design

a synthesis task and ask users to perform this task in the traditional way (see Section 1) and using the KG. In this setting, we could measure the required time, result quality and user satisfaction.

6 Preliminary Results

In this part, we conduct a preliminary experiment for table understanding using column headers. Here, we describe our hypothesis and the dataset we use. Then, we explain the experimental pipeline and discuss our initial results.

6.1 Hypothesis

We aim at understanding tabular datasets by inferring the schema of a corresponding KG using column headers. Our two experiments rely on this hypothesis: Interesting concepts in column headers could be captured using a clustering technique, such that cluster names nominate graph concepts while members show the related objects. For example, a cluster named *Author* has a member set { *Name, Email* } . This yields into two triples: *(Author, name, "Name")* and *(Author, email, "Email")*. However, manual user intervention is required to refine the resultant clusters. In fact, we cannot fully automate the conversion process from a CSV file into a KG [7, 22].

6.2 Dataset

We used a dataset [19] used in the compilation of data for the sWorm in our two experiments. This dataset presents information about earthworms in different geographical sites during a range of years. Additionally, it provides information on site level, species level, and metadata about the dataset itself. Our experiments shown here directly use the the column headers with at least one meaningful word. For example, *bio10_1* and *bio10_4* are excluded.

6.3 Experimental Pipeline

Figure 3 explains the pipeline of our experiments. The basic idea here is to apply a cosine distance-based clustering technique on the table headers represented in meaningful vectors. In fact, sWorm dataset has no unified convention of the headers; some are camel others are snake cases. So, the first block contains a parser which receives a list of headers, and it aims at getting a set of words inside the human-made header. The second component converts a header into a vector representation. We support two choices of vectors, either an ASCII code for the letters inside the header (syntactic representation) or using the word embeddings [15] (semantic representation). In this way, we can compute distances to determine the similarity among headers. After that, a distance-based clustering technique populates the initial clusters. Distance threshold would vary based on the type of vectors used. Then, the user has a facility to merge the clusters or to move some members from one cluster to another. The next component can suggest a cluster name based on the commonality among its members. If no common word found, *Unknown* would be the nominated name. Lastly, the user can rename the suggested names manually and export the schema in RDF/XML format.

6.4 Experimental Results & Discussion

Figure 4 illustrates the cosine distances among words' representation. Such that Equation 1 shows cosine distance between two vectors A, B. While Equation 2 gives the distance. We choose cosine similarity because it is independent on vector size, two vectors might be far apart by other metrics due to their sizes. As shown, the ASCII representation of the headers (Syntactic representation), is not a good discriminator among header names. Due to sharing a large amount of the same characters as in Figure

Table 1: Summary of experimental results

Representation	Granularity	No. Init. Clusters	Mistakes	Distance Threshold	Vector Dim.
Syntactic	Coarse-grained	4	14	0.15	82
Semantic	Fine-grained	11	6	0.6	1200

4a. Thus, it yields into a few but large (coarse-grained) clusters. However, the use of the pre-trained word embeddings (Semantic representation) discriminates among the headers very well, as in Figure 4b. We can conclude that the semantic representation is better than the syntactic representation in terms of the misplaced members (number of mistakes). But, it requires long vectors, such that a 300D vector represents each word. Although, the longest header consists of 4 words, so the final vector length for each header is a 1200D. Unlike the syntactic representation, which efficiently represents the header. Table 1 summarizes the results. We calculate the number of mistakes by comparing the initial clusters result against a manually created graph schema for the sWorm dataset. Thus, the more mistakes we have, the more user input we will need. In summary, this method would work well if we have relatively descriptive column headers that contain meaningful words. But, due to its constraints, this approach demonstrates the idea and requires additional information from table cells.

$$similarity = cos(\theta) = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

$$distance = 1 - similarity \tag{2}$$

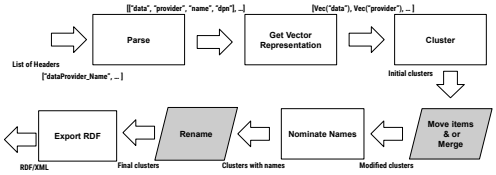
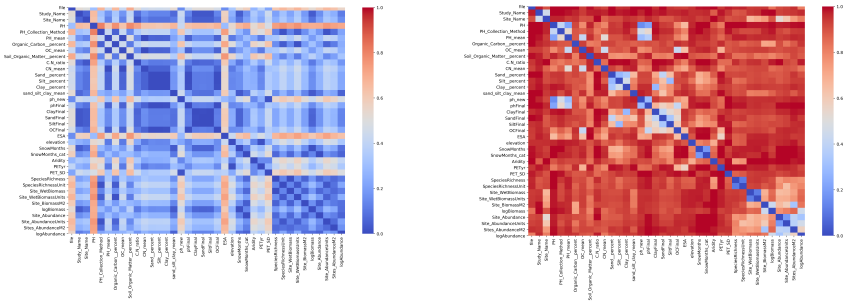


Fig. 3: Table header processing pipeline



(a) Syntactic distances (b) Semantic distances

Fig. 4: Distances using two different representations of the column headers, blue cells mean two words are close, red ones indicate large distance.

7 Conclusions

In this paper, we presented a KG construction framework. It processes various data sources initially from the Biodiversity domain. Mainly the tabular dataset itself, meta-

data, and the related publication. Our framework consists of three core modules: i) Concept Prediction, ii) Relation Detection, and iii) KG Construction which integrates the other two modules. Besides, we have discussed preliminary experiments[6] concerning table understanding using the column headers. Our results showed that the use of semantic embeddings as a column header representation is better than the syntactic one. Meanwhile, we will extend our existing methods to overcome the current limitations by considering column cells with headers. Moreover, we plan to make our proposed framework publicly available.

Acknowledgment

The authors thank the Carl Zeiss Foundation for the financial support of the project "A Virtual Werkstatt for Digitization in the Sciences (P5)" within the scope of the program line Breakthroughs: Exploring Intelligent Systems for Digitization - explore the basics, use applications. I thank Birgitta König-Ries, Joachim Denzler, and Sheeba Samuel for their guidance and feedback.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007). https://doi.org/10.1007/978-3-540-76298-0_52
2. Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M.E.: Towards a knowledge graph for science. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. pp. 1–6 (2018). https://doi.org/10.1145/3227609.3227689
3. Bach, N., Badaskar, S.: A review of relation extraction. Literature review for Language and Statistics II **2**, 1–15 (2007)
4. Batista-Navarro, R., Zerva, C., Ananiadou, S.: Construction of a biodiversity knowledge repository using a text mining-based framework. In: SIMBig. pp. 22–25 (2016)
5. Chabot, Y., Labbe, T., Liu, J., Troncy, R.: DAGOBAH: An end-to-end context-free tabular data semantic annotation system. CEUR Workshop Proceedings, vol. 2553, pp. 41–48. CEUR-WS.org (2019)
6. Chen, J., Jiménez-Ruiz, E., Horrocks, I., Sutton, C.: Colnet: Embedding the semantics of web tables for column type prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 29–36 (2019). https://doi.org/10.1609/aaai.v33i01.330129
7. Ermilov, I., Auer, S., Stadler, C.: User-driven semantic mapping of tabular data. In: Proceedings of the 9th International Conference on Semantic Systems. pp. 105–112. ACM (2013). https://doi.org/10.1145/2506182.2506196
8. Hassanzadeh, O., Efthymiou, V., Chen, J., Jimnez-Ruiz, E., Srinivas, K.: SemTab2019: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching - 2019 Data Sets (Oct 2019). https://doi.org/10.5281/zenodo.3518539
9. Kazama, J., Torisawa, K.: Exploiting wikipedia as external knowledge for named entity recognition. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). pp. 698–707 (2007)

[6] Code and the manually created KG are publicly available: `https://github.com/fusion-jena/ClusteringTableHeaders`

10. Knoblock, C.A., Szekely, P., Ambite, J.L., Gupta, S., Goel, A., Muslea, M., Lerman, K., Mallick, P.: Interactively mapping data sources into the semantic web. In: Proceedings of the First International Conference on Linked Science-Volume 783. pp. 13–24. CEUR-WS. org (2011)
11. Lehmberg, O., Ritze, D., Meusel, R., Bizer, C.: A large public corpus of web tables containing time and context metadata. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 75–76. International World Wide Web Conferences Steering Committee (2016). https://doi.org/10.1145/2872518.2889386
12. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. arXiv preprint arXiv:1812.09449 (2018)
13. Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and searching web tables using entities, types and relationships. Proceedings of the VLDB Endowment **3**(1-2), 1338–1347 (2010). https://doi.org/10.14778/1920841.1921005
14. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. arXiv preprint arXiv:1808.09602 (2018)
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
16. Nguyen, P., Kertkeidkachorn, N., Ichise, R., Takeda, H.: Mtab: Matching tabular data to knowledge graph using probability models. CEUR Workshop Proceedings, vol. 2553, pp. 7–14. CEUR-WS.org (2019)
17. Nishida, K., Sadamitsu, K., Higashinaka, R., Matsuo, Y.: Understanding the semantic structures of tables with a hybrid deep neural network architecture. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
18. Page, R.: Towards a biodiversity knowledge graph. Research Ideas and Outcomes **2** (2016). https://doi.org/10.3897/rio.2.e8767
19. Phillips, H.R., Guerra, C.A., Bartz, M.L., Briones, M.J., Brown, G., Crowther, T.W., Ferlian, O., Gongalsky, K.B., Van Den Hoogen, J., Krebs, J., et al.: Global distribution of earthworm diversity. Science **366**(6464), 480–485 (2019). https://doi.org/10.1101/587394
20. Smirnova, A., Cudré-Mauroux, P.: Relation extraction using distant supervision: A survey. ACM Computing Surveys (CSUR) **51**(5), 106 (2018). https://doi.org/10.1145/3241741
21. Thawani, A., Hu, M., Hu, E., Zafar, H., Divvala, N.T., Singh, A., Qasemi, E., Szekely, P.A., Pujara, J.: Entity linking to knowledge graphs to infer column types and properties. CEUR Workshop Proceedings, vol. 2553, pp. 25–32. CEUR-WS.org (2019)
22. Vander Sande, M., De Vocht, L., Van Deursen, D., Mannens, E., Van de Walle, R.: Lightweight transformation of tabular open data to rdf. In: I-SEMANTICS (Posters & Demos). pp. 38–42. Citeseer (2012)
23. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016). https://doi.org/10.1038/sdata.2016.18
24. Yao, L., Liu, H., Liu, Y., Li, X., Anwar, M.W.: Biomedical named entity recognition based on deep neutral network. Int. J. Hybrid Inf. Technol **8**(8), 279–288 (2015). https://doi.org/10.14257/ijhit.2015.8.8.29