On the Role of Preprocessing on Matching Tables to Knowledge Graphs

Vishvapalsinhji Parmar[1], Achraf Haddar[1] and Alsayed Algergawy[1]

[1]Chair of Data and Knowledge Engineering, University of Passau Passau, Germany

Abstract

Matching tabular data to knowledge graphs plays a crucial role in various applications, including Named Entity Recognition (NER). Data preprocessing has consistently shown to enhance the performance of data-driven systems. To this end, in this paper, we present a systematic preprocessing pipeline designed to improve the accuracy of table-to-graph matching by identifying and addressing anomalies in datasets from diverse domains, such as biodiversity, food, and Wikidata. Our pipeline, implemented over three iterations, focuses on correcting domain-specific irregularities to enhance data quality. Experimental results demonstrate substantial improvements, with F1 score increases of up to 50% in the food domain and 5% in biodiversity, surpassing existing methods. These advancements contribute to more efficient data interpretation and analysis across a variety of sectors.

Keywords

Table Annotation, Table Understanding, Preprocessing

1. Introduction

The exponential growth of online information offers substantial opportunities across various domains. However, the data is often in diverse and fragmented formats, categorized into structured, semi-structured, and unstructured forms. Among these, tabular data is widely used due to its readability and compactness, serving applications in fields such as medicine, climate change, biodiversity, and manufacturing [1]. Despite its prevalence, extracting meaningful information from tabular data remains challenging due to limited context, necessitating alignment with semantic artifacts like ontologies and knowledge graphs [2]. There are also some Knowledge graphs, which are networks of interconnected entities and concepts, provide a robust foundation for advanced data interpretation by leveraging relational information to uncover insights, support decision-making, and enhance analytics [3]. Aligning tabular data with knowledge graphs is thus a promising avenue for innovation. The Semantic Table Understanding (SemTab) Challenge has fostered advancements in algorithms that link tabular data to knowledge graphs, enhancing data comprehension[1]. A study in this field demonstrated that leveraging preprocessed cells using the Being Search API[2] can enhance system performance [4]. While many systems addressing this challenge focus on candidate generation, table element processing, and disambiguation [2], the diverse potential of the preprocessing stage remains underexplored.

To bridge this gap, we introduce a preprocessing pipeline designed to enhance data quality for table-to-knowledge graph matching. We carried out a number of experiments using datasets from Wikidata, Biodiversity, and Food Tables from the SemTab challenge. The results show notable improvements in Cell Entity Annotation (CEA) performance. The implementation code and notebooks for reproducibility are available in our repository[3].

EKAW 2024: EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands.

vishvapalsinhji.parmar@uni-passau.de (V. Parmar); alsayed.algergawy@uni-passau.de (A. Algergawy)

0000-0002-4370-2729 (V. Parmar); 0000-0002-8550-4720 (A. Algergawy)

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

[2]https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

[3]https://github.com/DKEPassau/PreprocessMatch

2. Motivation

Data preprocessing is a critical stage in matching tables to knowledge graphs, yet it is often overlooked. Although tabular data is inherently structured, it frequently contains noise, missing values, and inconsistencies that impede accurate interpretation. Effective preprocessing—comprising data cleaning, normalization, feature extraction, and transformation—enhances data quality, ensuring that it is ready for further processing and matching tasks. Our aim is to explore how preprocessing can significantly impact the performance of aligning tables to knowledge graphs. This study investigates various preprocessing techniques, such as handling missing values, language inconsistencies, and special characters, to improve data quality. We also examine the effectiveness of these techniques in the context of the Semantic Table Understanding (SemTab) challenge, where the primary goal is to achieve better annotation accuracy through systematic preprocessing.

By developing and implementing a structured preprocessing pipeline, we address specific challenges associated with different datasets. This study demonstrates how targeted preprocessing can enhance table annotation results, reduce errors, and ultimately improve the integration of tabular data with knowledge graphs, paving the way for more accurate and insightful data analysis.

3. Dataset and Exploratory Data Analysis (EDA)

To evaluate the effectiveness of our preprocessing pipeline, we employed three distinct datasets from the SemTab challenge: Wikidata Tables, BioDiversity Tables, and tFood Tables. Each dataset presents unique characteristics and challenges, requiring a tailored approach to preprocessing. Conducting a thorough Exploratory Data Analysis (EDA) was a critical first step in understanding these datasets, identifying their specific structures, and determining the necessary preprocessing steps to improve annotation accuracy.

3.1. Wikidata Tables

The Wikidata Tables dataset consists of synthetic data generated from the Wikidata knowledge graph using SPARQL queries. This dataset spans a wide range of domains and contains 500 tables, each stored as a separate CSV file. Each table includes a primary column designated as the subject, with additional columns providing contextual information. EDA revealed several data quality issues, including:

- **Missing Values:** Approximately 43.6% of the tables contain missing values, with 751 instances identified across the dataset. Since these values were not annotated in the original data, no imputation was performed.
- **Language Inconsistencies:** About 32.28% of the cell values are in non-English languages, identified using the 'Detect Language API.' This tool was chosen over the 'langid' Python package due to its superior accuracy in identifying languages like English, French, German, Spanish, Italian, and Arabic.
- **Special Characters and Misspellings:** Various cells contain special characters or spelling errors (e.g., 'City of Porsmouth' instead of 'City of Portsmouth'), which could hinder the retrieval of correct annotations. These issues were systematically addressed through data cleaning rules in the preprocessing pipeline.

3.2. BioDiversity Tables

The BioDiversity Tables dataset comprises 50 tables derived from real-world biodiversity research and manually annotated samples. It leverages three public repositories: data.world, BEFChina, and BExIS, and is characterized by four unique features: Specimen data, Numerical data, Abbreviations, and Special formats. EDA findings for this dataset include:

- **Numerical and String Data Types:** Over 54% of the columns contain numerical data, while 33% are strings. Most tables (49 out of 50) feature at least four columns, except for one single-column table.
- **Missing Values:** Similar to Wikidata, this dataset shows a high number of missing values (39,198), which also lack corresponding annotations in the CEA target file, indicating no need for imputation.
- **Domain-Specific Patterns:** The dataset contains domain-specific characteristics, such as species name abbreviations (e.g., 'C. sclerophylla' for 'Castanopsis sclerophylla') and composed values combining multiple elements. Tools like the NCBI Taxonomy database and ChatGPT were used to interpret these domain-specific nuances, ensuring accurate data preprocessing.

3.3. tFood Tables

The tFood dataset is designed specifically for the Food domain, including two types of tables: Horizontal Relational Tables and Entity Tables. Each table contains two columns, with one column representing entity properties (e.g., Prop0, Prop1) and the other providing detailed descriptions. Key findings from the EDA include:

- **String Data Type:** All cells earmarked for annotation are of the string type, requiring consistent handling of language and special characters.
- **Language Detection:** Using the 'Detect Language API,' it was identified that 11.92% of the annotated cells contain non-English content, necessitating appropriate preprocessing rules to manage multilingual data.
- **Composed Values and Special Cases:** Some columns feature values composed of multiple elements separated by characters like hyphens ('-'). Identifying these patterns was crucial to developing targeted preprocessing rules to ensure proper annotation.

3.4. Common Anomalies

In the analysis of tabular datasets, such as those from Wikidata, BioDiversity, and tFood, several common anomalies were identified that can hinder the accuracy of matching tables to knowledge graphs. These anomalies include:

- **Missing Values:** Frequently occurring in datasets, missing values can lead to incomplete or inaccurate data interpretation. For example, in the Wikidata and BioDiversity tables, a significant percentage of tables exhibited missing values, necessitating careful handling to avoid bias or errors in subsequent annotations.
- **Language Inconsistencies:** Datasets often contain content in multiple languages, complicating the retrieval of correct entities. For instance, over 32% of the Wikidata tables and nearly 12% of the tFood tables included non-English content. The use of the 'Detect Language API' helps in identifying and standardizing these inconsistencies.
- **Special Characters and Misspellings:** The presence of special characters (e.g., punctuation marks or symbols) and misspellings can mislead APIs or annotation tools, resulting in empty or incorrect query results. Correcting such errors is critical for improving data accuracy.
- **Cell Duplications and Composed Values:** Duplicated data entries or values composed of multiple elements (e.g., 'North-Lincolnshire') can create ambiguities in data interpretation. Developing rules to handle these cases ensures more accurate alignment of tabular data with knowledge graphs.
- **Abbreviations and Special Cases:** Domain-specific abbreviations (e.g., 'C. sclerophylla') require context-aware processing to ensure accurate annotation. Failure to correctly interpret these cases may result in significant annotation errors.

Addressing these anomalies is essential for maintaining the integrity of the data annotation process. To ensure a consistent and effective approach to preprocessing, we developed a systematic pipeline for anomaly detection and resolution. This pipeline is designed to apply standardized preprocessing steps across different datasets, reducing human intervention and enhancing overall annotation accuracy.

4. Pipeline

To effectively address the identified anomalies, we propose a robust pipeline for preprocessing and annotating cell values in tabular datasets. The pipeline aims to enhance the efficiency and accuracy of table annotation tasks by systematically addressing data quality issues. The proposed pipeline unfolds in three main phases, each targeting specific types of abnormalities identified during EDA. Figure 1 provides an overview of the sequential steps involved in the pipeline.

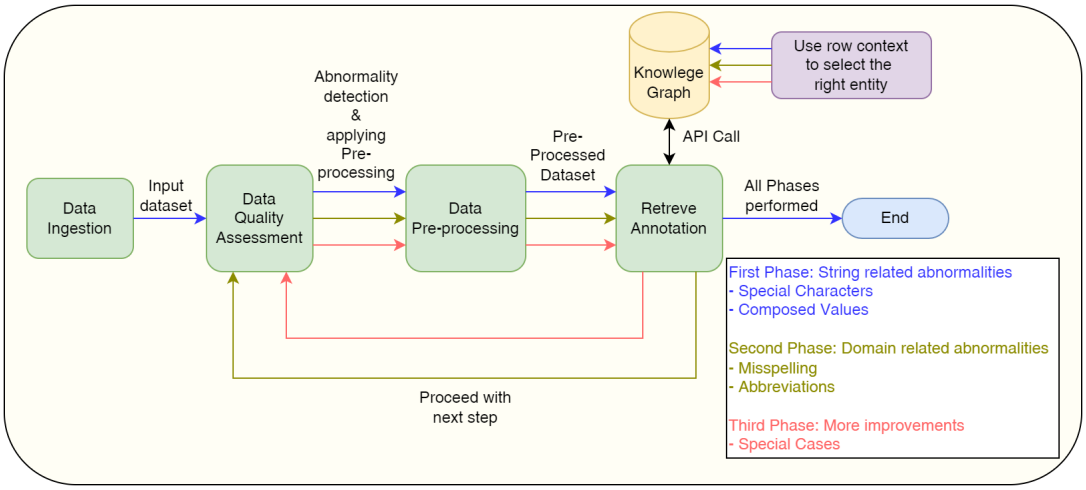

Figure 1: Proposed pipeline for table annotation

Phase 1: Detection of Obvious Anomalies : The first phase focuses on detecting and resolving obvious abnormalities in the dataset, such as special characters, missing values, and composed values separated by delimiters (e.g., hyphens). By applying specific preprocessing rules, we aim to standardize these elements to reduce ambiguity and improve the quality of the data being annotated.

Phase 2: Identification of Domain-Related Anomalies : The second phase addresses more complex anomalies, such as language inconsistencies, misspellings, and domain-specific abbreviations. This involves using tools like the 'Detect Language API' to identify non-English content and implementing targeted rules to correct misspellings and interpret abbreviations accurately. The goal is to reduce the number of incorrect or empty annotations.

Phase 3: Refinement and Advanced Preprocessing : The final phase involves refining the annotations further by exploring special cases and developing additional preprocessing rules to handle complex data patterns. This phase includes a comprehensive analysis of special characters, abbreviations, and composed values that could not be fully resolved in the earlier phases. The focus is on achieving the highest possible annotation accuracy by applying a refined set of rules tailored to the unique characteristics of each dataset.

Annotation Retrieval Process : After preprocessing, the refined dataset is ready for the annotation retrieval process. This involves querying external knowledge graphs to retrieve accurate annotations for each cell value. The process leverages contextual information from rows to select the correct entity from potential matches, ensuring a high degree of precision in the final annotations.

5. Experiments and Future Direction

To evaluate the performance of our preprocessing pipeline, we conducted experiments on three datasets: Wikidata, BioDiversity, and tFood. Each dataset underwent the proposed preprocessing steps, which were designed to address specific data anomalies such as missing values, language inconsistencies, special characters, and domain-specific patterns. The primary objective was to assess the impact of preprocessing on the Cell Entity Annotation (CEA) task accuracy. The F1 score was used as the main evaluation metric, representing the harmonic mean of Precision and Recall, to measure the effectiveness of our pipeline. Significant improvements were observed across all datasets, validating the robustness of our approach. The results showed an increase in F1 scores of 5% up to 50% in some cases, demonstrating the pipeline's effectiveness in enhancing annotation accuracy. Table 1 presents a summary of the F1 scores before and after applying the preprocessing pipeline, highlighting the improvements achieved.

Table 1
F1 score before and after applying pipeline

Dataset	F1 Score		Previous System Score
	Before	After	F1 Score
Wikidata Tables	0.681	0.845	0.88
BioDiversity Tables	0.385	**0.696**	0.64
tFood Tables	0.821	**0.858**	0.23[4]

The experimental results confirmed that targeted preprocessing steps, such as language detection, handling of special characters, and domain-specific adjustments, can significantly improve F1 scores. Our approach reduces errors and increases the reliability of data integration with knowledge graphs, providing a strong foundation for further development in this field. Future work may explore the optimization and automation of preprocessing pipelines and their application to additional domains. Integrating advanced machine learning techniques could further enhance anomaly detection and resolution, unlocking new possibilities for extracting insights from tabular data and advancing Semantic Table Understanding. Previous work, such as DREIFLUSS [5], highlights the potential to leverage preprocessed data, which can ultimately enhance system performance.

References

[1] R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, Information Fusion 81 (2022) 84–90.

[2] J. Liu, Y. Chabot, R. Troncy, V. Huynh, T. Labbé, P. Monnin, From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods, J. Web Semant. 76 (2023) 100761. doi:10.1016/j.websem.2022.100761.

[3] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3447772. doi:10.1145/3447772.

[4] E. G. Henriksen, A. M. Khorsid, E. Nielsen, A. M. Stück, A. S. Sørensen, O. Pelgrin, Semtex: A hybrid approach for semantic table interpretation, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, SemTab 2023, ISWC 2023, Athens, Greece, November 6-10, 2023, volume 3557 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 38–49. URL: https://ceur-ws.org/Vol-3557/paper3.pdf.

[5] V. R. Parmar, A. Algergawy, DREIFLUSS: A minimalist approach for table matching, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, SemTab 2023, ISWC 2023, Athens, Greece, November 6-10, 2023, volume 3557 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 50–60. URL: https://ceur-ws.org/Vol-3557/paper4.pdf.