Results of GRAMS+ at SemTab 2024

Binh Vu[1,*], Craig A. Knoblock[1] and Fandel Lin[1]

[1]USC Information Sciences Institute, Marina del Rey, CA 90292, USA

Abstract

There is an enormous number of tables available on the Web. However, it is difficult to automatically use the tables in data analytic pipelines because of the lack of semantic understanding of their structure and meaning. To address this problem, our approach, GRAMS+, automatically creates semantic descriptions of tables using distant supervision. SemTab is an annual challenge that provides a diverse set of benchmarks for systems that match tabular data with knowledge graphs. In this paper, we present the results of GRAMS+ at SemTab 2024 in the Accuracy Track. The results show that GRAMS+ is scalable and achieves competitive performance in the tasks in which we participated.

Keywords

SemTab 2024, Semantic Description, Semantic Table Interpretation, Knowledge Graphs, Semantic Web, Data Integration

1. Introduction

Matching tabular data to an ontology or a knowledge graph is an essential problem in Data Integration. The task is to annotate types of columns in the tables using classes of the target ontology and relations between columns using the ontology properties. We developed a novel approach, GRAMS+ [1], addressing this problem using distant supervision. The approach leverages the fact that some data in a table will often overlap with data in a knowledge graph (KG), which can be used to discover candidate types and relationships in the table. Then, the approach uses two neural networks (NN) trained with a labeled dataset generated automatically from Wikipedia tables to predict the final column types and relationships.

The Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) is an annual challenge with the goal of providing benchmarks and evaluations of existing solutions to this problem. In this paper, we present the results of GRAMS+ at the SemTab 2024 challenge focusing on the Accuracy Track. Our approach successfully annotates a very large number of tables and achieves first place on the tasks in which we participated.

2. The SemTab Challenge

The SemTab 2024 challenge consists of several tracks ranging from semantic table interpretation to dataset assessment and contributions. We focus on the Accuracy Track, which is relevant to our approach. This track contains four matching tasks: (1) the Cell Entity Annotation (CEA) matches a cell to a KG entity, (2) the Column Type Annotation (CTA) assigns a KG class to a column, (3) the Column Property Annotation (CPA) assigns a KG property to the relationship between two columns, and (4) Topic Detection (TD) assigns a KG class to a table. Figure 1 shows an example table annotation.

There are two types of tables in this track: horizontal tables (or relational tables) and entity tables. A horizontal table is a grid where each row represents an entity and each column shares the same semantic type (e.g., Figure 1). An entity table describes a single entity, where each row contains a property of that entity.

SemTab'24: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching 2024, co-located with the 23rd International Semantic Web Conference (ISWC), November 11-15, 2024, Baltimore, USA

binhvu@isi.edu (B. Vu); knoblock@isi.edu (C. A. Knoblock); fandel.lin@usc.edu (F. Lin)

https://binh-vu.github.io/ (B. Vu)

0000-0001-5808-9288 (B. Vu); 0000-0002-6371-4807 (C. A. Knoblock); 0000-0001-7024-2476 (F. Lin)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3889/paper6.pdf

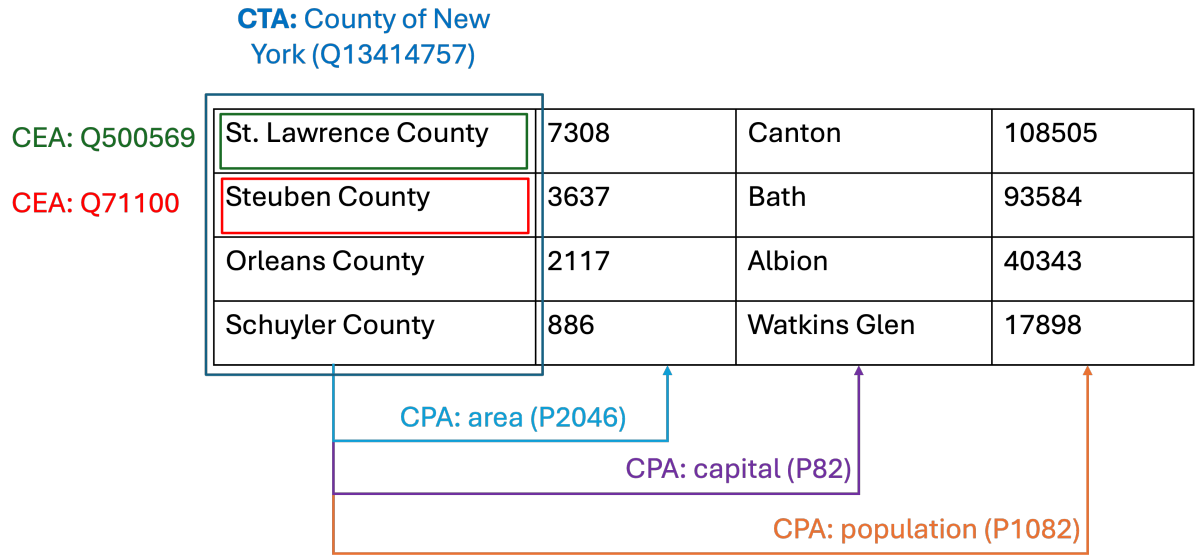

Figure 1: An example of a table with annotation

Finally, the standard micro precision, recall, and F_1 are used to measure the performance of the participating systems [2].

3. GRAMS+ Approach

Figure 2 shows the overall approach of GRAMS+. It starts by finding KG entities that are mentioned in a table. Then, we use a neural network (NN) to compute the scores of candidate entities of each table cell. The NN model is trained with a labeled dataset automatically generated from Wikipedia tables. Using the discovered candidates and their scores, we predict column types (CTA) and column relationships (CPA).

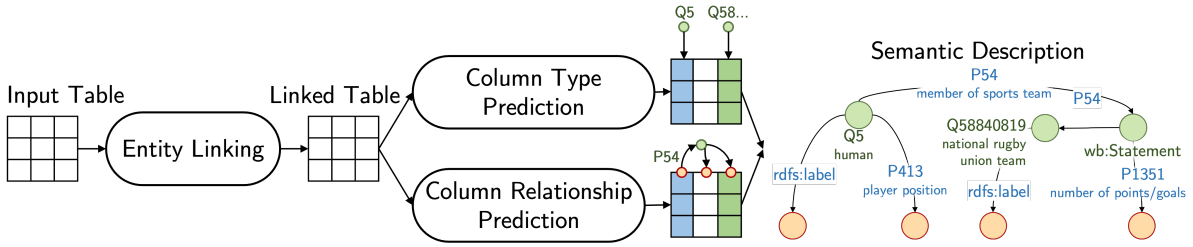

Figure 2: Overall approach of GRAMS+

We generate the labeled dataset by leveraging the hyperlinks inside the Wikipedia tables to find corresponding Wikidata entities and predict columns' relation- ships based on the linked entities. We remove context-inconsistent hyperlinks by first automatically assigning a type to each column based on the most common type of its entities. Then, we employ a blocklist to remove all links in a column if the column header is incompatible with the predicted column types. The blocklist is constructed by manually verifying headers that appeared in multiple predicted types. As our approach is detailed in [1], the remainder of this section provides a brief overview of each component in GRAMS+, along with any changes to fit the SemTab 2024 challenge.

3.1. Entity Linking

Following typical entity linking (EL) systems, our EL approach consists of three main steps: (1) detect the entity columns, which are the cells that will be linked; (2) retrieve candidate entities for each cell; and (3) compute the candidates' likelihood.

For step 1, we directly use the target entity columns provided in SemTab's datasets instead of running the entity detection. To retrieve candidate entities, GRAMS+ combines multiple search strategies such as using public Wikidata Search API, keyword search using ElasticSearch, and fuzzy search using SymSpell. Given the huge number of tables in the Wikidata Tables dataset in Round 2 (78,745 tables), we cannot use the public Wikidata API to search and only use the two later strategies.

To compute the candidates' likelihood, we use a two-hidden-layer perceptron with RELU activations. It is trained using the auto-label dataset with the following groups of features:

Surface Features include four string similarity functions between a cell and an entity name: Levenshtein, Jaro-Winkler, Monge Elkan, and Generic Jaccard.

Entity-context Similarity Features capture the coherence between a candidate and the surrounding context of a cell. GRAMS+ uses two context similarity features: the weighted dot product of the column header and the candidate description, and the number of cells matched with the candidate's property divided by a large constant representing the maximum number of columns in a table (e.g., 20) for rescaling. The embeddings are computed from a Sentence Transformer model [3][1], and the weights of embedding dimensions are learnable parameters. Note that GRAMS+ trains two entity linking models for tables with and without headers. Because tables from the SemTab datasets do not have column headers, GRAMS+ uses the model trained on tables without headers.

Entity Prior Features bias the predictions toward popular entities. Currently, we use the normalized log page rank of a candidate as the prior feature. The normalized log page rank of an entity e is calculated as follows:

$$\frac{\log(\text{pagerank}(e)) - \min_{e' \in \mathscr{E}} \log(\text{pagerank}(e'))}{\max_{e' \in \mathscr{E}} \log(\text{pagerank}(e')) - \min_{e' \in \mathscr{E}} \log(\text{pagerank}(e'))}$$

where $\mathscr{E}$ is the set of entities in KG, pagerank(e) is the pagerank of an entity e.

3.2. Column Type Prediction

To predict the type of a column, we use a greedy algorithm that first selects the type with the highest score from the set of types directly found in the candidate entities of a column. Then, it iteratively refines the prediction by replacing it with an ancestor type within d distance of the directed types if the score difference is larger than a specific threshold δ until d reaches the maximum chosen distance (max_distance). The score of a type is computed by summing the maximum likelihood of the candidate entities of the type for each cell and then dividing by the number of rows. We use the same threshold (δ = 0.1) and maximum distance (max_distance = 2) as in [1].

3.3. Column Relationship Prediction

To predict the relationship of a column, GRAMS+ first constructs a candidate graph containing potential relationships between columns. Then, GRAMS+ uses a classifier to predict the likelihood of each link in the graph. As the SemTab challenge provides pairs of target columns for predictions, we directly use the most likely relationships between target columns as the final predictions.

The classifier employed to predict the likelihood of links is also a two-hidden-layer perceptron with RELU activations. It is trained on the auto-label dataset with features such as the relative frequency of discovering the link from top K entities, the average link likelihood, the relative frequency of finding contradicting information between the table data and KG data, and whether there is a many-to-many relationship between the source and target of the link.

[1]We use the pretrained all-mpnet-base-v2 model.

Table 1
Performance of GRAMS+ on CPA and CTA tasks. Precision and F_1 scores are reported in percentage

Dataset	CPA				CTA			
	F_1	Precision	Recall	Rank	F_1	Precision	Recall	Rank
Wikidata Tables round 1	89.8	98.8	82.30	1	92.9	92.9	92.9	1
Wikidata Tables round 2	89.9	99.2	82.19	1	95.6	95.6	95.6	1

4. SemTab 2024 Results

Table 1 reports the performance of GRAMS+ on the Wikidata Tables datasets. We cannot run GRAMS+ on the tBiodiv and tBiomed datasets because the values of the subject columns', which contain the main entities, were anonymized. Since the names are changed, these datasets focus on a different aspect of the problem, which is identifying the anonymous entities. This is not the focus of GRAMS+, and we leave it for future work.

At the time of writing the paper, GRAMS+ achieves first place among the participants on the Wikidata Tables datasets. The two datasets, in total, have approximately 109,000 tables. This shows that GRAMS+ is scalable and can handle a large number of tables.

5. Related Work

Table Understanding is an essential problem in Data Integration and has attracted many studies over the years. A comprehensive related work to GRAMS+ can be found in [1]. In this section, we briefly discuss work related to GRAMS+ in the setting of the SemTab challenge.

Most systems participating in the SemTab, including GRAMS+, exploit the existing knowledge in a KG. Typically, they first identify KG entities in a table (CEA) and match the properties of entities with values in the table to find column types (CTA) and relationships between columns (CPA). The best performing systems in SemTab such as MTab [4], DAGOBAH [5], and others such as KGCode-Tab [6], LinkingPark [7], BBW [8], TorchicTab-Heuristic [9], and SemTex [10] improve various aspects of the pipeline such as candidate entity retrieval, scoring functions to rank the matched results, or repeat the pipeline several times or until reaching equilibrium. Compared to GRAMS+, they often rely on hand-crafted scoring functions, while GRAMS+ uses distant supervision to learn to classify correct entities and column relationships. Moreover, GRAMS+ tackles a general setting where we need n-ary relationships to correctly model data in the tables.

The SemTab 2023 and 2024 also include other tasks, such as Table Topic Detection and Matching Table Metadata to KG. These are not the focus problems of GRAMS+, and we leave them for future work.

6. Conclusion

This paper presents the results of GRAMS+, a distant supervised approach for annotating column types and relationships of tables, for the SemTab 2024 Accuracy Track. GRAMS+ achieves rank 1 for datasets on which it was evaluated.

In future work, we plan to improve the performance of GRAMS+ by jointly predicting column types and relationships. We also plan to extend GRAMS+ to leverage table context, metadata, and modeling instructions to support tables without overlapping data to a target knowledge graph.

Acknowledgements

This material is based upon research supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112390132 and Contract No. 140D0423C0093. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA); or its Contracting Agent, the U.S. Department of the Interior, Interior Business Center, Acquisition Services Directorate, Division V.

References

[1] B. Vu, C. A. Knoblock, B. Shbita, F. Lin, Exploiting distant supervision to learn semantic descriptions of tables with overlapping data, in: The Semantic Web–ISWC 2024: 23th International Semantic Web Conference, ISWC 2024, November 11–15, 2024, Proceedings 20, Springer International Publishing, 2024.

[2] O. Hassanzadeh, N. Abdelmageed, V. Efthymiou, J. Chen, V. Cutrona, M. Hulsebos, E. Jiménez-Ruiz, A. Khatiwada, K. Korini, B. Kruit, et al., Results of semtab 2023, in: CEUR Workshop Proceedings, volume 3557, 2023, pp. 1–14.

[3] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-Networks (2019). arXiv:1908.10084.

[4] P. Nguyen, I. Yamada, N. Kertkeidkachorn, R. Ichise, H. Takeda, SemTab 2021: Tabular data annotation with MTab tool, http://ceur-ws.org/Vol-3103/paper8.pdf, ???? Accessed: 2023-10-6.

[5] V.-P. Huynh, Y. Chabot, T. Labbé, J. Liu, R. Troncy, From heuristics to language models: A journey through the universe of semantic table interpretation with DAGOBAH, Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2022).

[6] X. Li, S. Wang, W. Zhou, G. Zhang, C. Jiang, T. Hong, P. Wang, KGCODE-Tab results for SemTab 2022, https://ceur-ws.org/Vol-3320/paper5.pdf, ???? Accessed: 2023-10-6.

[7] S. Chen, A. Karaoglu, C. Negreanu, T. Ma, J.-G. Yao, J. Williams, A. Gordon, C.-Y. Lin, LinkingPark: An integrated approach for semantic table interpretation, http://ceur-ws.org/Vol-2775/paper7.pdf, ???? Accessed: 2023-10-6.

[8] R. Shigapov, P. Zumstein, J. Kamlah, L. Oberlander, J. Mechnich, I. Schumm, bbw: Matching CSV to wikidata via meta-lookup, https://madoc.bib.uni-mannheim.de/57386/3/paper2.pdf, ???? Accessed: 2023-10-6.

[9] I. Dasoulas, D. Yang, X. Duan, A. Dimou, TorchicTab: Semantic Table Annotation with Wikidata and Language Models, in: CEUR Workshop Proceedings, 2023, pp. 21–37.

[10] E. G. Henriksen, A. M. Khorsid, E. Nielsen, A. M. Stück, A. S. Sørensen, O. Pelgrin, Semtex: A hybrid approach for semantic table interpretation, 2023.