Embedding-based data matching for disparate data sources

Nour Elhouda Kired, Franck Ravat, Jiefu Song, Olivier Teste

Embedding-based data matching for disparate data sources

Nour Elhouda Kired[1,2][0009−0009−0610−6280], Franck Ravat[1][0000−0003−4820−841X], Jiefu Song[1][0000−0002−2066−7051], and Olivier Teste[2][0000−0003−0338−9886]

[1] Université Toulouse Capitole, IRIT, Toulouse France
[2] Université Toulouse II Jean Jaurès, IRIT, Toulouse France
nour-elhouda.kired@irit.fr

Abstract. Dealing with heterogeneous sources is an important challenge in the field of knowledge discovery and management. Schema matching methods are employed to solve this problem using three approaches: schema-based, instance-based, or a combination. This paper focuses on mapping between a schema-available (only) data source and a data source containing both schema and instance (both). Given the lack of suitable methods for aligning these two types of sources, we propose an approach using embedding models to provide vector modelling of sources and calculate similarities between data. Our solution consists in combining domain-specific embedding models and cross-domain embedding models to make data matching possible and efficient between the above-mentioned data sources. We have conducted several experiments using the Valentine datasets to evaluate our data matching method on several disparate tabular data. The result indicate effectiveness in terms of stability and ablation handling.

Keywords: Schema Matching · Disparate Data Source · Embeddings.

1 Context & Main Issues

Data lakes store diverse data types, making schema-on-read essential for determining data schema during access. Aligning these disparate sources at the schema level is crucial for comprehensive data analysis, employing schema-based, instance-based, or hybrid methods. Integrating unstructured data like images and videos requires metadata, posing a challenge when matching sources with schemas and instances to those with only schemas. Existing methods, including rule-based techniques like Regular Expressions, often fall short due to disparate data structures and semantic complexities, despite improvements from incorporating syntactic and semantic relations [1]. This paper proposes a novel approach combining cross-domain and domain-specific embeddings to effectively match non-comparable data source structures, preserving the structure of the input sources.

In schema matching, pre-trained embedding models, particularly cross-domain embedding models, have shown effectiveness in generalizing knowledge across

different domains. Zhang et al. [2] and Dash et al. [3] use BERT-based embeddings to perform schema matching, leveraging deep domain knowledge. The BART model [4] excels in Natural Language Inference (NLI) for understanding complex semantic relationships, essential for schema matching [5]. Cappuzzo et al. [6] use domain-specific embeddings tailored to the characteristics of specific domains, employing graph embedding to uncover connections within datasets and enhance pattern matching accuracy. While cross-domain embeddings generalize knowledge across domains without considering data structure, domain-specific embeddings focus on relationships within structured data but struggle with disparate data sources. Recent research in Open Domain Question Answering (ODQA) and information retrieval uses a dual-model approach, fine-tuning BERT-based models for domain-specific data retrieval [7], but these methods are not fully adapted for schema matching with varied data structures.

This paper contributes by developing an approach to align schemas from disparate data source structures, focusing on mapping schema-based data with both schema and instance-level data (e.g., CSVs, Excels) in Data Lakes. Our method combines cross-domain embeddings, capturing broad semantic relationships, with domain-specific embeddings, using graph embeddings for deeper dataset insights. This approach efficiently aligns disparate data structures by leveraging both broad and deep domain knowledge.

2 Proposed Framework

2.1 Problem Statement

We present an approach using the BART model for cross-domain embeddings and EMBDI for domain-specific embeddings to map two disparate data sources: $S1$ (schema $Sc1$) and $S2$ (schema $Sc2$ and instances $I2$). The goal is to align $Sc1$ and $Sc2$, producing matches $(a_i, b_j, score)$ that indicate the degree of similarity between elements from the two schemas, with scores ranging from 0 to 1.

2.2 Overview

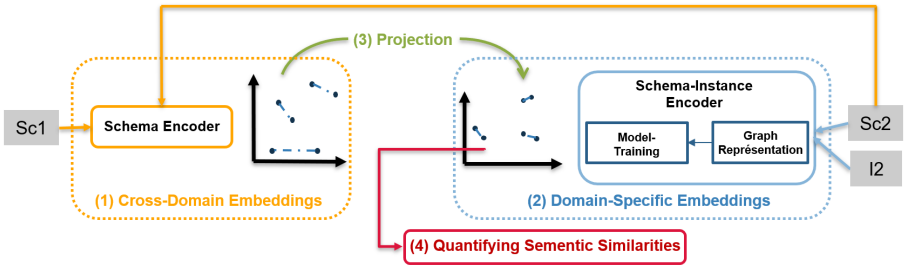

Fig. 1: Framework Overview

Our schema-matching approach with four key components significantly contributing to achieving our goal. Figure 1 illustrates our approach with these components:

(1) Cross-Domain Embedding. We employ the BART model's Natural Language Inference (NLI) capabilities to compare the semantic and syntactic relationships between two schemas, Sc_1 and Sc_2. Each attribute in Sc_2 is treated as a hypothesis against the attributes in Sc_1. The BART model outputs logits for these attribute pairs, which are then transformed into probabilities to form a similarity matrix textitSimMatrix. Let: A = $\{a_1, a_2, \ldots, a_m\}$ be the set of attributes in schema Sc_1, B = $\{b_1, b_2, \ldots, b_n\}$ be the set of attributes in schema Sc_2. For each pair (a_i, b_j), the BART model computes logits l_{ij}, which are converted to probabilities p_{ij} representing the likelihood of entailment: p_{ij} = softmax(l_{ij})

The similarity matrix *SimMatrix* is then defined as:

$$\textbf{SimMatrix}_{ij} = p_{ij} \quad \text{for} \quad i = 1, \ldots, m \quad \text{and} \quad j = 1, \ldots, n$$

(2) Domain-Specific Embedding. The EmbDI algorithm constructs domain-specific embeddings by representing relational data as a heterogeneous graph $G = (V, E)$. Nodes V include token nodes (T), Record ID nodes (RID), and Column ID nodes (CID). Edges E represent relationships between these nodes. Random walks on G generate sequences of nodes, forming sentences for the embedding training corpus. The sentences are used to train embeddings with algorithms like word2vec. Let $\mathbf{E}_G(v)$ be the embedding vector for node $v \in V$.

(3) Projection of Sc_1 Attributes onto Domain-Specific Embeddings. Attributes from Sc_1 are projected onto the domain-specific embeddings using the similarity matrix *SimMatrix*.

For each attribute $a_i \in A_1$, the embedding vector $\mathbf{E}_{Sc_1}(a_i)$ is computed as:

$$\mathbf{E}_{Sc_1}(a_i) = \sum_{j=1}^{n} \textbf{SimMatrix}_{ij} \cdot \mathbf{E}_{Sc_2}(b_j)$$

Where : $\mathbf{E}_{Sc_1}(a_i)$ is the embedding vector of attribute $a_i \in Sc_1$ and $\mathbf{E}_{Sc_2}(b_j)$ is the embedding vector of attribute $b_j \in Sc_2$.

(4) Quantifying Semantic Similarities. To quantify the semantic similarities between attributes in Sc_1 and Sc_2, we use cosine similarity. For each pair (a_i, b_j):

$$\text{similarity}(a_i, b_j) = \frac{\mathbf{E}_{Sc_1}(a_i) \cdot \mathbf{E}_{Sc_2}(b_j)}{\|\mathbf{E}_{Sc_1}(a_i)\| \|\mathbf{E}_{Sc_2}(b_j)\|}$$

This results in a list of matches (a_i, b_j, score) where $a_i \in Sc_1$ and $b_j \in Sc_2$, with each match assigned a similarity score.

3 Experiments

In this section, we describe experiments to evaluate the effectiveness of our schema matching approach, focusing on two main research questions:

- **RQ1(Effectiveness & stability)** How significantly does the proposed schema mapping approach enhance the F1 score when aligning data schemas of disparate structures? And how consistent is our method when the configuration and dataset are held constant?
- **RQ2(Ablation)** How does the removal of a specific phase of the proposed schema matching approach affect its overall performance?

We address these questions through in-depth experiments, exploring various configurations and parameters to refine our approach to schema matching. In our experiments, we used four dataset categories from the "Valentine" collection [8]. The experimental setup utilized PyTorch with NVIDIA CUDA on 12 Dell servers with Intel Xeon processors. Comprehensive results and additional experiments (e.g., handling noise) are available on the companion website.[3].

3.1 RQ1. Effectiveness & stability

Table 1: Performance Metrics Across Dataset Categories and Relation Types

Dataset categories	**Relations (Datasets)**	**F1 Score**		**Recall**		**Precision**	
		Mean	Std	Mean	Std	Mean	Std
ChEMBL (180 datasets)	Joinable (48)	0.92	0.08	0.95	0.09	0.91	0.11
	Semantically-Joinable (48)	0.95	0.07	0.98	0.06	0.93	0.11
	Unionable (36)	0.98	0.02	0.95	0.05	1.00	0.00
	View-Unionable (48)	0.96	0.05	0.98	0.04	0.94	0.08
TPC-DI (180 datasets)	Joinable (48)	0.76	0.22	0.97	0.06	0.67	0.25
	Semantically-Joinable (48)	0.74	0.21	0.96	0.06	0.65	0.26
	Unionable (36)	0.98	0.02	0.97	0.03	1.00	0.00
	View-Unionable (48)	0.75	0.21	0.95	0.06	0.67	0.25
Wikidata (4 datasets)	Joinable (1)	0.67	0.01	0.67	0.02	0.67	0.00
	Semantically-Joinable (1)	0.93	0.04	0.87	0.08	1.00	0.00
	Unionable (1)	0.87	0.04	0.77	0.06	1.00	0.00
	View-Unionable (1)	0.66	0.02	0.67	0.00	0.66	0.04
Magellan	Unionable (7)	0.93	0.06	0.88	0.11	1.00	0.00

Discussion: The results in Table 1, using 96 configurations, demonstrate the effectiveness and reliability of our schema matching algorithm with an average accuracy of 0.85 across all datasets and relations. Low standard deviations in ChEMBL (0.055), Wikidata (0.022), and Magellan (0.06) indicate robustness. However, variability in TPC-DI (0.75-0.98) and some Wikidata instances (0.66 for joinable and view-unionable) highlights challenges in matching semantically

[3] https://github.com/user28060/Embedding-based-data-matching-for-disparate-data-sources.git

similar but differently labeled attributes, especially when these attributes have high similarity scores within the same domain.

A detailed stability analysis was conducted on datasets D1-D10, each tested with different hyperparameter configurations. We focused on three key hyperparameters affecting random walks, resulting in 12 configurations, each run 10 times. The results show high stability, with D1-D7 achieving mean F1, Recall, and Precision scores of 1.00. The lower F1 score for D10 (0.91) is due to difficulties distinguishing between 'givenNameLabel' and 'familyName.' Slight variations in standard deviation were observed for D8 (0.01), D9 (0.002), and D10 (0.018), confirming the algorithm's consistent performance across different configurations.

3.2 RQ2. Ablation

The ablation study evaluated two approaches: the combined approach and the cross-domain approach. A domain-specific model alone was unfeasible due to the incompatibility of heterogeneous dataset structures with the EMBDI algorithm. Experiments were conducted on various datasets with noise. The study used 96 configurations to compare the performance of the two methods, with results shown in Figure 2.

Discussion: The combined approach outperforms the cross-domain approach with a higher median F1 score and tighter interquartile range, demonstrating better and more consistent performance. Figure 2.b indicates that the combined approach is, on average, 0.3 F1 score points better than the cross-domain approach, making it preferable for applications where performance consistency is crucial.

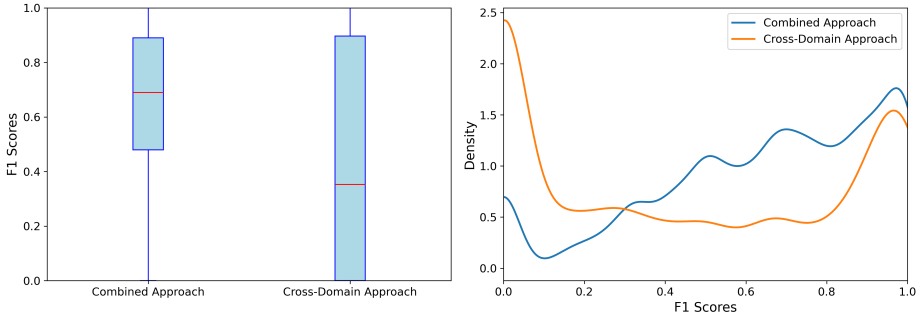

Fig. 2: Comparison of Cross-Domain Embeddings and Combined Approaches

4 Conclusion

Mapping disparate data source structures is critical within data lakes. This paper introduces a novel schema matching method to align schema-based data

sources with those containing both schemas and instances. By leveraging embedding models, we created vector representations to compute data similarities. Our technique, merging domain-specific and cross-domain models, was rigorously evaluated using Valentine dataset categories. Results highlight its effectiveness, stability, and proficiency in ablation, indicating its potential in data matching.

Future work will focus on enhancing our framework by incorporating entity resolution and developing diverse models for cross-domain and domain-specific contexts, aiming to improve both schema matching and entity resolution.

References

1. Christodoulou, K., Fernandes, A.A.A., Paton, N.W.: Combining Syntactic and Semantic Evidence for Improving Matching over Linked Data Sources. In: Wang, J., Cellary, W., Wang, D., Wang, H., Chen, S.-C., Li, T., and Zhang, Y. (eds.) Web Information Systems Engineering – WISE 2015. pp. 200–215. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-26190-4_14.
2. Zhang, Y., Floratou, A., Cahoon, J., Krishnan, S., Müller, A.C., Banda, D., Psallidas, F., Patel, J.M.: Schema Matching using Pre-Trained Language Models. In: 2023 IEEE 39th International Conference on Data Engineering (ICDE). pp. 1558–1571. IEEE, Anaheim, CA, USA (2023). https://doi.org/10.1109/ICDE55515.2023.00123.
3. Dash, S., Bagchi, S., Mihindukulasooriya, N., Gliozzo, A.: Linking Tabular Columns to Unseen Ontologies. In: Payne, T.R., Presutti, V., Qi, G., Poveda-Villalón, M., Stoilos, G., Hollink, L., Kaoudi, Z., Cheng, G., and Li, J. (eds.) The Semantic Web – ISWC 2023. pp. 502–521. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-47240-4_27.
4. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, http://arxiv.org/abs/1910.13461, (2019).
5. Liu, H., Cui, L., Liu, J., Zhang, Y.: Natural Language Inference in Context - Investigating Contextual Reasoning over Long Texts. AAAI. 35, 13388–13396 (2021). https://doi.org/10.1609/aaai.v35i15.17580.
6. Cappuzzo, R., Papotti, P., Thirumuruganathan, S.: Creating Embeddings of Heterogeneous Relational Datasets for Data matching Tasks. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. pp. 1335–1349. ACM, Portland OR USA (2020). https://doi.org/10.1145/3318464.3389742.
7. 9. Bosch, N., Shalmashi, S., Yaghoubi, F., Holm, H., Gaim, F., Payberah, A.H.: Fine-Tuning BERT-based Language Models for Duplicate Trouble Report Retrieval. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 4737–4745. IEEE, Osaka, Japan (2022). https://doi.org/10.1109/BigData55660.2022.10020825.
8. Koutras, C., Siachamis, G., Ionescu, A., Psarakis, K., Brons, J., Fragkoulis, M., Lofi, C., Bonifati, A., Katsifodimos, A.: Valentine: Evaluating Matching Techniques for Dataset Discovery, http://arxiv.org/abs/2010.07386, (2021).