DB Strange: Exploring the Multiverse of Entity Resolution Datasets

Keywords: entity resolution, data matching, benchmarking, dataset versions

Benchmarking Entity Resolution

Entity resolution (ER) is an essential step in data cleaning pipelines. It aims to detect and consolidate multiple records that refer to the same real-world entity. This topic has been studied in the literature for more than 50 years, but many challenges remain. One of them is the **lack of consolidated benchmarks** for evaluating and comparing ER approaches. This lack is exacerbated by the fact that state-of-the-art ER approaches are based on supervised learning methods, which are particularly data-hungry. More specifically: (i) There is no centralized repository of ER benchmarks. Rather, they are fragmented across multiple websites (e.g., [1, 2, 3, 4, 5], to list a few). (ii) **Different versions** of the same benchmark exist, so comparing multiple approaches requires selecting the correct dataset versions and understand how these versions were created by transforming the original data.

paper-id	#records	#attributes	#cluster	#dupl.pairs	largest cluster	comment	source	year
6	1,295	-	112	-	-	-	McCallum	2005
12	1,875	12	-	17,184	-	-	McCallum	2019
13	1,879	-	182	-	238	corrected labels	McCallum	2019
14	1,295	-	-	17,184	-	-	SecondString	2011
15	-	-	-	-	-	-	HPI	2013
20	1,799	-	-	-	-	-	SecondString	2018
21	1,3k	12	-	17k	-	-	-	2019
24	-	-	-	-	-	XML	McCallum	2009
27	2,191	5	305	-	-	-	McCallum	2006
28	1,916	-	121	-	-	-	-	2000
37	1,295	-	122	-	-	-	-	2003
40	1,878	-	191	-	236	-	-	2016
41	1,295	-	-	17,184	-	-	-	2016
49	1,295	12	116	-	-	corrected labels	-	2007
52	-	9	-	-	-	-	-	2020
55	1,838	5	190	-	-	-	McCallum	2014

Figure 1: Use of CORA dataset across papers

For instance, the CORA dataset has been extensively used over the last decades. As illustrated in Figure 1, the reported statistics about this dataset differ significantly, implying the existence of different dataset versions. This limits the comparability of the results of different ER studies.

Project Goal

We will follow an entire research cycle from problem inception and literature research to algorithm development and, finally, to evaluation and deployment of a publicly available ER-dataset repository. Together, we will prepare a research article and submit it to an international conference.

Our goal is to bring clarity to the embarrassingly diffuse landscape of ER-Benchmark datasets. We lay a special focus on the analysis of different dataset versions. We want to develop algorithms that can (i) cluster ER-datasets to identify different versions of the same data and (ii) characterize the differences of multiple dataset versions using frameworks such as Explain-

Da-V [5]. We start the project with a literature search phase. Afterwards, we will collect ER-dataset versions typically used to benchmark ER-algorithms. Here we are not starting from scratch, but already have a few dataset versions as a starting point. Then, we will design and develop our algorithms for clustering and characterizing those dataset versions. Thereafter, we will conduct a sensitivity analysis and check whether using different versions of the same dataset has an impact on the reported quality of several state-of-the-art ER-approaches, such as Ditto [6]. Finally, we will build our repository and develop a dashboard to help users navigate within this repository.

In summary, our approach consists of the following tasks:

- Collect available ER benchmark datasets
- Provide datasets and ground truths in a unified format
- Store meta-information about the datasets
- Cluster datasets to identify different versions of the same data
- Characterize the differences of multiple dataset versions
- Measure impact of different dataset versions (sensitivity analysis)
- Create a dashboard that allows filtering and querying the dataset repository
- Rights management (are we allowed to share them?)

Experience in the following is beneficial:

- Entity resolution and data analysis
- Proficiency in Python and JavaScript

Programs:

- IT-Systems Engineering MA
- Data Engineering MA
- Software Systems Engineering MA

Contact Details

This project will be supervised by Lukas Laskowski, Dr. Fabian Panse, and Prof. Dr. Felix Naumann at the Information Systems chair. If you have any questions, please do not hesitate to contact us.

References

[1] https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md

[2] https://webdatacommons.org/largescaleproductcorpus/wdc-products/

[3] https://data.dws.informatik.uni-mannheim.de/benchmarkmatchingtasks/index.html

[4] https://sites.google.com/site/anhaidgroup/useful-stuff/the-magellan-data-repository?authuser=0

[5] https://dbs.uni-leipzig.de/research/projects/benchmark-datasets-for-entity-resolution

[6] Roee Shraga, Renee Miller: *Explaining Dataset Changes for Semantic Data Versioning with Explain-Da-V*, Proc. VLDB Endow. 16(6): 1587-1600, 2023.

[7] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, Wang-Chiew Tan: *Deep Entity Matching with Pre-Trained Language Models*. Proc. VLDB Endow. 14(1): 50-60, 2020.