Towards a Controllable and Realistic Entity Matching Benchmark

Xue Li[1*], Zeyu Zhang[2*], Sebastian Schelter[3], Paul Groth[2], Madelon Hulsebos[1]

[1]CWI, Netherlands [2]UvA, Netherlands [3]TU Berlin, Germany

effy.li@cwi.nl, z.zhang2@uva.nl

* Equal contribution.

Entity Matching (EM), the task of identifying records that refer to the same real-world entities across heterogeneous data sources, is a fundamental component of data preparation. However, existing EM benchmarks [1, 2] suffer from three limitations that obstruct robust model evaluation. First, they are often static and inflexible, lacking the ability to adapt to different domains, or allow for controllable manipulation of data disturbances. Second, they often fail to reflect real-world complexity, especially in an enterprise setting [3], for example the use of domain-specific codenames or semantically challenging aliases that are difficult for rule-based methods to handle. Third, modern EM techniques, which are increasingly reliant on large language models (LLMs) [4], face a significant data contamination risk, as their massive pre-training corpora likely include the entities used in existing benchmarks.

To address these limitations and facilitate more robust evaluation, we present EMBench, a controllable and realistic EM benchmark. EMBench contains a data generation pipeline that includes four modules: (i) source data preparation; (ii) synthetic data generation; (iii) syntactic and semantic variation generation; (iv) entity pair generation. This benchmark allows users to generate datasets with varying domains and levels of complexity in a flexible manner, enabling the evaluation of EM systems under diverse conditions. Furthermore, EMBench reflects real-world scenarios by incorporating elements such as domain-specific codenames (e.g., qids such as Q38111 for Leonardo DiCaprio) and semantically challenging perturbations (e.g., "Rasheed Ferrache" and "Aaron Heff" are referring to the same person). Crucially, to mitigate potential data leakage, EMBench utilizes a synthetic data generation process to create novel, counterfactual entities that are unlikely to be present in LLMs' training data, while preserving the functional dependencies implied in the original schema.

In this work, we demonstrate EMBench for three domains: movies, sports events and music records. These datasets contain a total of 45,708 entities, resulting in 228,540 candidate pairs with 40% positive matches. We evaluate existing EM systems on EMBench and surface weaknesses to data perturbations that have previously not been uncovered. For example, we find a drop of up to 68% in performance with a combination of synthetic and semantic variations. We envision that EMBench will enable robustness testing of existing and novel EM systems to ensure better generalizability across domains and use-cases.

References

[1] Konda, P., et al. (2016). Magellan: toward building entity matching management systems. PVLDB, 9(12), 1197–1208.

[2] Peeters, R., et al. (2024). WDC Products: A multi-dimensional entity matching benchmark. Proc. EDBT 2024, 22–33.

[3] Bodensohn, J.-M., et al. (2024). LLMs for data engineering on enterprise data. VLDB TaDA Workshop 2024.

[4] Narayan, A., et al. (2022). Can foundation models wrangle your data? PVLDB, 16(4), 738–746.