Knowledge Transfer for Entity Resolution with Siamese Neural Networks

MICHAEL LOSTER, IOANNIS KOUMARELAS, and FELIX NAUMANN,
Hasso Plattner Institute, University of Potsdam, Germany

The integration of multiple data sources is a common problem in a large variety of applications. Traditionally, handcrafted similarity measures are used to discover, merge, and integrate multiple representations of the same entity—duplicates—into a large homogeneous collection of data. Often, these similarity measures do not cope well with the heterogeneity of the underlying dataset. In addition, domain experts are needed to manually design and configure such measures, which is both time-consuming and requires extensive domain expertise.

We propose a deep Siamese neural network, capable of learning a similarity measure that is tailored to the characteristics of a particular dataset. With the properties of deep learning methods, we are able to eliminate the manual feature engineering process and thus considerably reduce the effort required for model construction. In addition, we show that it is possible to transfer knowledge acquired during the deduplication of one dataset to another, and thus significantly reduce the amount of data required to train a similarity measure. We evaluated our method on multiple datasets and compare our approach to state-of-the-art deduplication methods. Our approach outperforms competitors by up to +26 percent F-measure, depending on task and dataset. In addition, we show that knowledge transfer is not only feasible, but in our experiments led to an improvement in F-measure of up to +4.7 percent.

CCS Concepts: • **Information systems** → **Entity resolution**; **Deduplication**; • **Computing methodologies** → **Neural networks**; **Transfer learning**;

Additional Key Words and Phrases: Entity resolution, duplicate detection, transfer learning, neural networks, metric learning, similarity learning, data quality

ACM Reference format:
Michael Loster, Ioannis Koumarelas, and Felix Naumann. 2021. Knowledge Transfer for Entity Resolution with Siamese Neural Networks. *J. Data and Information Quality* 13, 1, Article 2 (January 2021), 25 pages.
https://doi.org/10.1145/3410157

1 DUPLICATE DETECTION

The need to integrate multiple data sources into a single dataset is present in many application areas. A major challenge that arises during the integration process emerges from the fact that records from different data sources often contain duplicate entries, i.e., several entries that refer to the same real-world entity. These duplicates, implying poor data quality, can directly affect

Authors' addresses: M. Loster, I. Koumarelas, and F. Naumann, Hasso Plattner Institute, University of Potsdam, Potsdam, Germany; emails: {michael.loster, ioannis.koumarelas, felix.naumann}@hpi.de.

1936-1955/2021/01-ART2 $15.00
https://doi.org/10.1145/3410157