(Artificial) Mind over Matter
Humans In and Humans Out in Matching

Roee Shraga
Supervised by Prof. Avigdor Gal and Prof. Rakefet Ackerman
Technion – Israel Institute of Technology, Haifa, Israel
shraga89@campus.technion.ac.il

ABSTRACT

The matching task is at the heart of data integration, in charge of aligning elements of data sources. Historically, matching problems were considered semi automated tasks in which correspondences are generated by matching algorithms and subsequently validated by human expert(s). This research is devoted to the changing role of humans in matching, which is divided into two main approaches, namely *Humans Out* and *Humans In*. With the increase in amount and size of matching tasks, the role of humans as validators seems to diminish; thus *Humans In* questions the inherent need for humans in the matching loop. On the other hand, *Humans Out* focuses on overcoming human cognitive biases via algorithmic assistance. Above all, we observe that matching requires unconventional thinking demonstrated by advance machine learning methods to complement (and possibly take over) the role of humans in matching.

1. INTRODUCTION

Modern industrial and business processes require intensive use of large-scale *data alignment and integration* techniques to combine data from multiple heterogeneous data sources into meaningful and valuable information. Data alignment and integration has been recently challenged by the need to handle large volumes of data, arriving at high velocity from a variety of sources, which demonstrate varying levels of veracity. This challenging setting, often referred to as *big data*, renders many of the existing techniques, especially those that are human-intensive, obsolete.

At the heart of the data integration realm lies the matching task [2], in charge of aligning elements of data sources. In particular, whenever data sources are represented as schemata, the task of *schema matching* aligns attributes that convey similar semantic content. At the data level, *entity resolution* (also known as *record deduplication*) aims at "cleaning" a database by identifying tuples representing the same entity. Initial heuristic attempts (*e.g.*, COMA [4]) were followed by theoretical grounding (*e.g.*, see [2, 6]), algorithmic solutions for efficient and effective integration, and a body of systems, benchmarks and competitions that allow comparative empirical analysis of integration solutions.

Matching problems have been historically defined as a semi-automated task in which correspondences are generated by matching algorithms and outcomes are subsequently validated by one or more human experts. The reason for that is the inherent assumption that humans "do it better." The traditional roles of humans and machines in matching are subject to change due to the availability of data and advances in machine learning. Therefore, in the proposed research we question this assumption and aim at developing a machine learning framework for matching.

Given the availability of data and the improvement of machine learning techniques, this line of research is devoted to the investigation of respective roles of humans and machines in achieving cognitive tasks in matching, aiming to determine whether traditional roles of humans and machines are subject to change [15, 16]. Such investigation, we believe, will pave a way to better utilize both human and machine resources in new and innovative manners. We consider two possible modes of change, namely *humans out* and *humans in*. *Humans Out* aim at exploring out-of-the-box latent matching reasoning using machine learning algorithms when attempting to overpower human matcher performance. Pursuing out-of-the-box thinking, we investigate the best way to include machine and deep learning in matching. *Humans in* explores how to better involve humans in the matching loop by assigning human matchers with a symmetric role to algorithmic matcher in the matching process.

In following sections we describe each of the two modes of change. Section 2 describes how and where we envision replacing humans in the matching loop. In Section 3, we detail our approch to better involve humans in matching by understanding their strengths and weaknesses. Finally, we summarize and discuss future directions in Section 4.

Proceedings of the VLDB 2020 PhD Workshop, August 31st, 2020. Tokyo, Japan.

2. HUMANS OUT

The *Humans Out* approach seeks matching subtasks, traditionally considered to require cognitive effort, in which humans can be excluded. An initial good place to start is with the basic task of identifying correspondences. We note that many contemporary matching algorithms use heuristics, where each heurisitc associates some semantic cue to justify an alignment between elements. For example, string-based matchers use string similarity as a cue for item alignment. We observe that such heuristics, in essence, encode human intuition about matching. Our earlier work [17] showed that human matching choices can be reasonably predicted by classifying them into types, where a type correspond to an existing heuristic. Moreover, in our experiments, decision making of most human matchers can be predicted well using a combination of two algorithmic matchers. Therefore, we can argue that the cognitive effort of many human matchers can be easily replaced with such heuristics.

Next, we describe two works aiming to enhance the automation of matching, focusing on the task of schema matching. The main component of these works is a *similarity matrix*, a conceptual model representing a matching result.

2.1 Learning to Rerank Schema Matches

In [7, 9] we suggested a learning algorithm for re-ranking top-K matches so that the best match is ranked at the top termed $LRSM$ (illustrated in Figure 1). The proposed algorithm has shown good results when tested on real-world as well as synthetic datasets, offering an alternative to humans in selecting the best match, a task traditionally reserved for human verifiers.

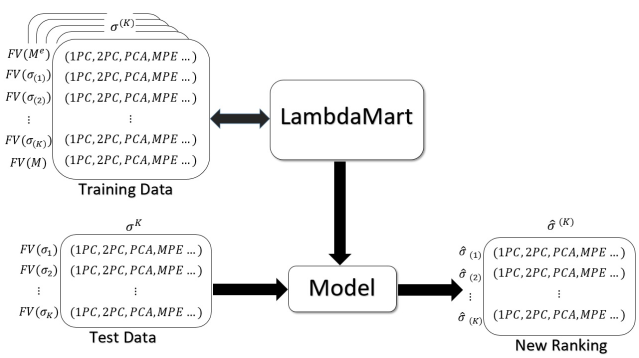

Figure 1: *Learn-to-Rerank Schema Matches* ($LRSM$) algorithm illustrated

The novelty of $LRSM$ is in the use of similarity matrices as a basis for learning features, creating feature-rich datasets that fit learning and provide us with a feature aggregation that is needed to enrich algorithmic matching beyond that of human matching. To create a reranking framework, we adopt a learning-to-rank approach [3], utilizing matching predictors [8, 13] as features. In addition to the state-of-the-art predictors, which mostly emphasize positive characteristics of a match, we propose a novel set of matching predictors that capture complementary negative aspects.

We show a bound on the size of K, given a desired level of confidence in finding the best match, justified theoretically and validated empirically. This bound is useful for top-K algorithms [11] and, as psychological literature suggests, also applicable when introducing a list of options (as in the traditional top-K setting) to humans [14].

Finally, using large scale experiments with real-world benchmark ontology and schema sets, as well as synthetic data, we show the effectiveness of the proposed algorithmic solution. Specifically, we show that the size of a top-K match list is geometrically distributed with a parameter that can be estimated as the amount of times the original best match was the one with the highest $F1$ value. Additionally, we show empirical evidence for the theoretical choice of K, demonstrate that the newly suggested predictors correlate well with evaluation measures, validate the use of NDCG as an optimization function, and above all show that $LRSM$ performs better than state-of-the-art methods providing improved (and robust) matching results.

2.2 Cross-Domain Schema Matching using Deep Similarity Matrix Adjustment and Evaluation

In a recent paper [18], we show that deep learning can also be applied to "small" matching problems such as schema matching, making extensive use of similarity matrices. We offer a novel post processing step to schema matching that improves the final matching outcome without human intervention. We present a new mechanism, *similarity matrix adjustment*, to calibrate a matching result and propose an algorithm (dubbed ADnEV) that manipulates, using deep neural networks, similarity matrices, created by state-of-the-art algorithmic matchers.

ADnEV uses *deep neural networks*, providing a data-driven approach for extracting hidden representative features for an automatic schema matching process, removing the requirement for manual feature engineering. ADnEV learns two conjoint neural network models for adjusting and evaluating a similarity matrix. ADnEV algorithm applies these models to iteratively adjust and evaluate new similarity matrices (illustrated in Figure 2), created by state-of-the-art matchers. With such a tool at hand, we enhance the ability to introduce new data sources to existing systems without the need to rely on either domain experts (knowledgeable of the domain but less so on the best matchers to use) or data integration specialists (who lack sufficient domain knowledge). Having a trained ADnEV model also supports systems where human final judgement is needed by regulation, *e.g.,* healthcare, by offering an improved matching recommendation.

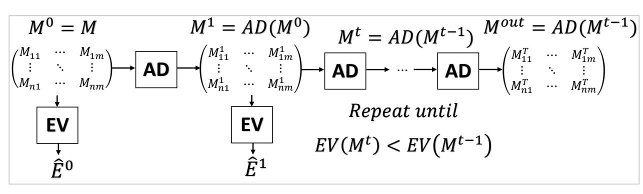

Figure 2: ADnEV algorithm illustrated

We empirically demonstrate the effectiveness of ADnEV for improving matching results, using real-world benchmark ontology and schema sets. We show that ADnEV can generalize into new domains without the need to learn the domain terminology, thus allowing cross-domain learning. We also show ADnEV to be a powerful tool in handling schemata which matching is particularly challenging. Finally, we show the benefit of using ADnEV in a related integration task of ontology alignment.

3. HUMANS IN

The *Humans In* approach aims at investigating whether the current role humans take in the matching process is effective and whether alternative role can improve overall performance of the matching process.

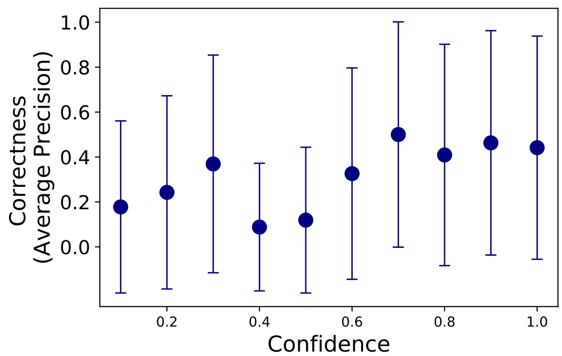

Figure 3: Correctness by confidence, partitioned into buckets of 0.1

By way of motivation, we provide an illustration (Figure 3) of the relationship between human confidence in matching and correctness (in terms of precision) based on our experiments [1, 17]. It is clear that human subjective confidence cannot serve as a good predictor to matching correctness. Next, we describe a work that shows how human biases affect confidence levels via consistency dimensions.

3.1 A Cognitive Model of Human Matching Bias

A recent study [1], aided by metacognitive models, analyzes the consistency of human matchers. We explore three main consistency dimensions as potential cognitive biases, taking into account the time it takes to reach a matching decision, the extent of agreement among human matchers and the assistance of algorithmic matchers. In particular, we showed that when an algorithmic suggestion is available, humans tend to accept it to be true, in sharp contradiction to the conventional validation role of human matchers.

Interestingly enough, all dimensions were found predictive of both confidence and accuracy of human matchers. This indicates that 1) humans have cognitive biases affecting their ability to provide consistent matching decisions, and 2) that such biases has predictive value in determining to what extent a human matcher's alignment decision is accurate. Our empirical evaluation serves as a proof-of-concept that validates the important roles of humans as participants in the matching process, and less so as validators. As an example, Figure 4 compares confidence with correctness, by showing the proportion of correctly identified correspondences, out of all correspondences (*i.e.,* precision), partitioned according to elapsed time (red) and mean of confidence across all human matchers, again partitioned according to elapsed time (blue). For each measure we also include a linear trend-line and error bars (standard deviation) for each time bucket.

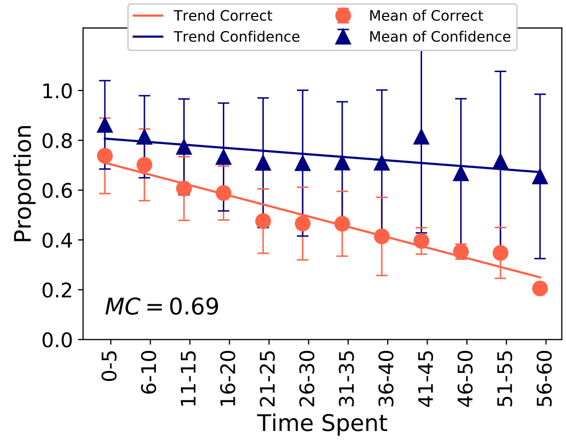

Figure 4: Temporal Dimension: Confidence (Blue) and correctness (Red) by elapsed time

As time passes, less decisions made by humans are correct and there is a decline of human confidence.

3.2 InCognitoMatch: Cognitive-aware Matching via Crowdsourcing

Acknowledging cognitive awareness in human matching, we recently proposed InCognitoMatch [19], the first cognitive-aware crowdsourcing application for matching tasks. InCognitoMatch provides a handy tool to validate, annotate, and correct correspondences using the crowd whilst accounting for human matching biases. In addition, InCognitoMatch enables system administrators to control context information visible for workers and analyze their performance accordingly. For crowd workers, InCognitoMatch is an easy-to-use application that may be accessed from multiple crowdsourcing platforms. In addition, workers completing a task are offered suggestions for followup sessions according to their performance in the current session. We foresee that such a tool will become handy in matching schemata in big data setting, where schema description may be poorly documented and human expertise becomes a scarce resource.

4. ONGOING AND FUTURE RESEARCH

In this paper we presented our approach for human involvement in the matching loop, introducing tasks where humans can be replaced and emphasizing our vision for understanding human behavior to allow better engagement. An additional overarching goal is to propose a common matching framework that would allow treating matching as a unified problem whether we match schemata attributes, ontology elements, process activities, entity's tuples, *etc.* Next, we describe some concrete ongoing and future research directions.

Cognition-aware Matching Collaboration: Match consistency was introduced in [1] as a measure of human matching variability along potential bias dimensions. As a direct future direction, we design a collaboration matcher that combines human and algorithmic opinions to improve the matching outcome by compensating for human biases along consistency dimensions as defined [1], namely temporal, consesnsuality, and control. We validated the proposed matcher using an empirical study with human and algorithmic matchers over a well-known benchmark, showing it provides better matching performance than human or algorithmic matching, performed separately.

Expert Identification: In [1] we show that humans have cognitive biases decreasing their ability to perform matching tasks effectively (see Section 3.1). *Expert identification* aims to predict humans qualification to serve as "experts" for a matching task. We intend to explore predictive behaviors that capture the process of human matching by transforming physical aspects (such as time, screen scrolls, mouse tracking, and eye movement) into features that can be used for examining the role of humans in the matching process. This, in turn, would enable matching systems to carefully select a matching expert that fits the task.

Learning from Matchers: Using machine learning for data integration raises the issue of shortage of labeled data to offer supervised learning [5, 9, 10, 12, 18]. Hence, pursuing less-than-supervised (*e.g.*, unsupervised, weakly supervised) methods would be a natural next step to follow. In a nutshell, we will propose a framework that uses pre-trained embeddings to represent data elements, processes a candidate pair to be matched with bidirectional LSTM, and trained using state-of-the-art heuristic matchers. Once trained, the framework will be independent of both human input and human designed heuristics. Initial empirical evaluation shows the proposed framework to performs better than multiple baselines and provide insights on future technique choices.

Matching Relevance: The vision we put forward is for the creation of a *probabilistic matching relevance* framework that will allow matching tasks to consider matching *intent* when creating a match. An intent reflects user preferences that may relate to granularity level, system requirement, match context, or simply individual inclination. We will present a probabilistic model of a match, showing that intent, either implicitly or explicitly specified, enables more accurate matching by better separating the relevant from the irrelevant. The proposed probabilistic notation will describe uncertainty in general existing matching problem, and accompanied with an intent, will enable assessment of the relevance of a match to a system rather than its correctness.

Acknowledgments

I would like thank Dr. Haggai Roitman, Dr. Tomer Sagi, Dr. Ofra Amir, and Coral Scharf, for their involvement in this research.

5. REFERENCES

[1] R. Ackerman, A. Gal, T. Sagi, and R. Shraga. A cognitive model of human bias in matching. In *PRICAI: Trends in Artificial Intelligence*, 2019.

[2] Z. Bellahsene, A. Bonifati, and E. Rahm, editors. *Schema Matching and Mapping*. Data-Centric Systems and Applications. Springer, 2011.

[3] C. J. Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11:23–581, 2010.

[4] H. H. Do and E. Rahm. Coma: a system for flexible combination of schema matching approaches. In *Proceedings of VLDB*, pages 610–621. VLDB Endowment, 2002.

[5] X. L. Dong and T. Rekatsinas. Data integration and machine learning: A natural synergy. In *SIGMOD*, pages 1645–1650. ACM, 2018.

[6] A. Gal. *Uncertain Schema Matching*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.

[7] A. Gal, H. Roitman, and S. Roee. Heterogeneous data integration by learning to rerank schema matches. In *IEEE International Conference on Data Mining, ICDM*. IEEE Computer Society, 2018.

[8] A. Gal, H. Roitman, and T. Sagi. From diversity-based prediction to better ontology & schema matching. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1145–1155. International World Wide Web Conferences Steering Committee, 2016.

[9] A. Gal, H. Roitman, and R. Shraga. Learning to rerank schema matches. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2019.

[10] F. Jabeen, H. Leopold, and H. A. Reijers. How to make process model matching work better? an analysis of current similarity measures. In *International Conference on Business Information Systems*, pages 181–193. Springer, 2017.

[11] C. Macdonald, R. L. T. Santos, and I. Ounis. The whens and hows of learning to rank for web search. *Information Retrieval*, 16(5):584–628, Oct 2013.

[12] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra. Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 International Conference on Management of Data*, pages 19–34. ACM, 2018.

[13] T. Sagi and A. Gal. Schema matching prediction with applications to data source discovery and dynamic ensembling. *The VLDB Journal*, 22(5):689–710, 2013.

[14] B. Schwartz. The paradox of choice: Why more is less. Ecco New York, 2004.

[15] R. Shraga. (artificial) mind over matter: Integrating humans and algorithms in solving matching problems. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD)*. ACM, 2018.

[16] R. Shraga and A. Gal. The changing roles of humans and algorithms in (process) matching. In *International Conference on Business Process Management*, pages 106–109. Springer, 2019.

[17] R. Shraga, A. Gal, and H. Roitman. What type of a matcher are you?: Coordination of human and algorithmic matchers. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2018, Houston, TX, USA, June 10, 2018*, pages 12:1–12:7, 2018.

[18] R. Shraga, A. Gal, and H. Roitman. Adnev: Cross-domain schema matching using deep similarity matrix adjustment and evaluation. *PVLDB*, 13(9):1401–1415, 2020.

[19] R. Shraga, C. Scharf, R. Ackerman, and A. Gal. Incognitomatch: Cognitive-aware matching via crowdsourcing. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD*. ACM, 2020.