CERTEM: Explaining and Debugging Black-box Entity Resolution Systems with CERTA

Tommaso Teofili
Roma Tre University, Red Hat
Rome, Italy
tommaso.teofili@uniroma3.it
tteofili@redhat.com

Donatella Firmani
Sapienza University
Rome, Italy
donatella.firmani@uniroma1.it

Nick Koudas
University of Toronto
Toronto, Canada
koudas@cs.toronto.edu

Paolo Merialdo
Roma Tre University
Rome, Italy
paolo.merialdo@uniroma3.it

Divesh Srivastava
AT&T Chief Data Office
New Jersey, USA
divesh@research.att.com

ABSTRACT

Entity resolution (ER) aims at identifying record pairs that refer to the same real-world entity. Recent works have focused on deep learning (DL) techniques, to solve this problem. While such works have brought tremendous enhancements in terms of effectiveness in solving the ER problem, understanding their matching predictions is still a challenge, because of the intrinsic opaqueness of DL based solutions. Interpreting and trusting the predictions made by ER systems is crucial for humans in order to employ such methods in decision making pipelines. We demonstrate CERTEM an explanation system for ER based on CERTA, a recently introduced explainability framework for ER, that is able to provide both saliency explanations, which associate each attribute with a saliency score, and counterfactual explanations, which provide examples of values that can flip a prediction. In this demonstration we will showcase how CERTEM can be effectively employed to better understand and debug the behavior of state-of-the-art DL based ER systems on data from publicly available ER benchmarks.

PVLDB Reference Format:
Tommaso Teofili, Donatella Firmani, Nick Koudas, Paolo Merialdo, and Divesh Srivastava. CERTEM: Explaining and Debugging Black-box Entity Resolution Systems with CERTA. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at URL_TO_YOUR_ARTIFACTS.

1 INTRODUCTION

Entity Resolution (ER) is the task that aims at matching records that refer to the same real-world entity. Although widely studied for the last 50 years in many research communities, ER still represents a challenging data management problem. Many works have investigated the application of machine learning (ML) techniques to solve the ER problem, more recently several works employing deep learning (DL) models have been shown to improve ontraditional approaches [3, 5, 7, 10]. Typically applying an ML approach to the ER problem involves the training of a classifier, possibly a deep neural network, for this problem. Given a set of training data, i.e. record pairs, and associated labels (match or non-match), a classifier is trained to solve a binary classification problem. The obtained classifier is then used to predict if any pair of records is to be considered a match or non-match.

Since DL models are typically considered black-box systems, recent researches focused on the exploration of techniques to offer *explanations* for predictions, aiming to reveal how the DL network reaches its decision [6]. Producing the explanations for an ER system has several important practical implications. For example, explanations can help understand the rationale behind an instance that is misclassified by the ER system; this way we can better plan interventions to fix such mistakes. Explanations can also help to check whether an ER system is making correct predictions for sound reasons.

Saliency and *counterfactual* methods [1] are the most commonly used explainability approaches. In the context of explaining the results of a classifier for ER, saliency methods aim at identifying the most influential attributes in an input pair, with respect to the predicted outcome. For this sake a *saliency* score is assigned to each record attribute. Counterfactual explanation methods help understand the behavior of the system by generating different ways the original input can be altered so that a different outcome is predicted by the ER system.

In this paper we demonstrate CERTEM: a tool based on CERTA [12][1] that provides saliency and counterfactual explanations for multiple ER systems. CERTEM also employs the generated explanations to detect biases and improve ER systems. CERTA is an explainability framework for ER systems.

The demonstration makes the following contributions:

- We provide insights on ER systems' behaviors through saliency and counterfactual explanations.
- Working on same input pairs, using CERTEM, we show how different DL based ER systems are influenced by different attributes.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit https://creativecommons.org/licenses/by-nc-nd/4.0/ to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

[1] preprint available at https://github.com/tteofili/certa/blob/master/preprint.pdf

- We demonstrate how CERTEM can be used to identify biases in training data using saliency explanations.
- We demonstrate how CERTEM can use the generated counterfactual explanations to augment training data and improve the behavior of the ER system for wrongly predicted samples.

2 BACKGROUND

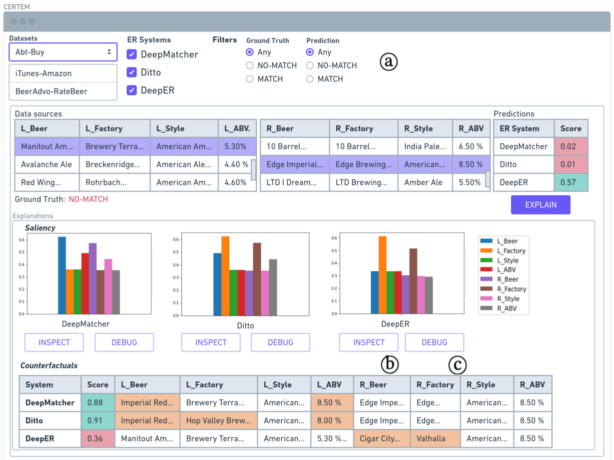
This section provides a brief overview of how CERTA generates saliency and counterfactual explanations for ER systems.

Open triangles. Consider the need to explain an ER system, trained on records $u \in U$ and $v \in V$, predicting an input pair $\langle u, v \rangle$ to be *non-matching* (resp. *matching*). CERTA builds what we call an *open triangle* over $\langle u, v \rangle$ by finding another record $w \in U$ (called *support record*) such that the same ER system predicts $\langle w, v \rangle$ to be *matching* (resp. *non-matching*). If we *perturb* u by progressively copying attribute values from w to u, deriving a u', increasingly making u' more similar to w based on their content, at some point the prediction of the model will flip, declaring u' and v to be a *match* (resp. *non-match*). In a left open triangle u and v are called respectively the *free* and *pivot* records and $w \in U$, viceversa in a right open triangle v is the *free record* while u is the *pivot record* and $w \in V$. Repeating the same procedure for many support records produces evidence of the influence that attributes and sets of attributes have on the input prediction.

Probabilities of necessity and sufficiency. CERTA defines the *saliency* of an attribute in the prediction outcome as the probability that changing the value of that attribute is a *necessary* factor for flipping the outcome of the prediction (*probability of necessity*), across different open triangles. Symmetrically, CERTA generates *counterfactual* explanations having the highest probability that changing the value of a certain *set of attributes* is a *sufficient* factor for flipping the outcome of a prediction (probability of sufficiency). To calculate such probabilities, CERTA uses a frequentist approach. The number of times an attribute is changed over the number of actual flips gives the *probability of necessity*, the number of times changing a set of attributes results in a flip over the number of times that the set of attributes is changed gives the *probability of sufficiency*.

Computing probabilities on lattices. CERTA associates a lattice structure to each generated open triangle. Every such lattice is built on the partial order between the elements of the power set of the attributes to be altered and the subset inclusion relation. Hence in a lattice built over a left (resp. right) open triangle, each node is associated with a specific set of attributes to be altered in the *free record*. Each node of the lattice is then *tagged* as 1 if copying the values of its corresponding attributes from the *support record* into the corresponding attributes of the *free record* leads to flipping the original output, 0 otherwise. Tagging each node is performed by visiting the lattice bottom-up with a breadth-first strategy. For each visited node, CERTA computes the prediction associated to the perturbation corresponding to the attributes of the node.

CERTA has two outputs: (i) the saliency of each attribute a is calculated as the probability of necessity of a and (ii) the counterfactual explanations corresponding to sets of attributes having the highest probability of sufficiency and lowest cardinality.



CERTEM

Datasets

Abt-Buy

iTunes-Amazon

BeerAdvo-RateBeer

ER Systems

☑ DeepMatcher

☑ Ditto

☑ DeepER

Filters

Ground Truth
● Any
○ NO-MATCH
○ MATCH

Prediction
● Any
○ NO-MATCH
○ MATCH

ⓐ

Data sources

L_Beer	L_Factory	L_Style	L_ABV.	R_Beer	R_Factory	R_Style	R_ABV
Manitout Am...	Brewery Terra...	American Am...	5.30%	10 Barrel...	10 Barrel...	India Pale...	6.50 %
Avalanche Ale	Breckenridge...	American Ale...	4.40 %	Edge Imperial...	Edge Brewing...	American...	8.50 %
Red Wing...	Rohrbach...	American Am...	4.60%	LTD I Dream...	LTD Brewing...	Amber Ale	5.50%

Predictions

ER System	Score
DeepMatcher	0.02
Ditto	0.01
DeepER	0.57

Ground Truth: NO-MATCH

EXPLAIN

Explanations

Saliency

Counterfactuals

System	Score	L_Beer	L_Factory	L_Style	L_ABV	R_Beer	R_Factory	R_Style	R_ABV
DeepMatcher	0.88	Imperial Red...	Brewery Terra...	American...	8.50 %	Edge Impe...	Edge...	American...	8.50 %
Ditto	0.91	Imperial Red...	Hop Valley Brew...	American...	8.00 %	Edge Impe...	Edge...	American...	8.50 %
DeepER	0.36	Manitout Am...	Brewery Terra...	American...	5.30 %...	Cigar City...	Valhalla	American...	8.50 %

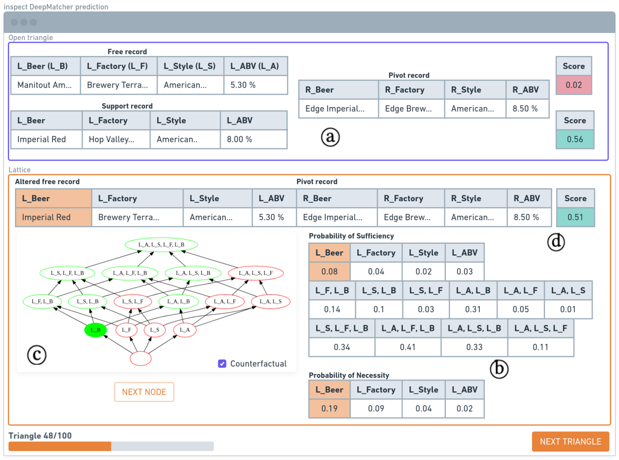
Figure 1: CERTEM explaining different ER systems

3 SYSTEM OVERVIEW

This section describes CERTEM, a system to generate saliency and counterfactual explanations for Entity Resolution systems based on the CERTA explanation framework described in Section 2. CERTEM makes it possible to generate, inspect and use saliency and counterfactual explanations on different ER systems and multiple datasets. Users select pairs of records to be explained by choosing among the available datasets or by referencing new ones. Users can also filter the available records by their ground truth label or ER system predicted outcome.

Upon selection of a pair of records (Figure 1 ⓐ), CERTEM can generate explanations for all the selected ER systems, providing a unified visual comparison of which portions of the input records were most important according to each different ER system for making their predictions. At this stage users can *drill down* the behavior of each explained system by either *inspecting* (Figure 1 ⓑ) the generated explanations or *debugging* (Figure 1 ⓒ) the system through the explanations. Inspecting an explanation drives users inside the inner workings of CERTA, showing how it generates open triangles (Figure 2 ⓐ) and how altering each attribute influences CERTA's probabilities of necessity and sufficiency (Figure 2 ⓑ). The UI of CERTEM shows a visual view of lattices associated to each open triangle (Figure 2 ⓒ) and gives users the possibility of going through CERTA open triangles one by one and interactively generating explanations in a step by step way.

For explanations to be actionable, they need to be faithful to an ER system, therefore CERTEM allows users to visualize the faithfulness of a saliency explanation to an ER system by showing how altering the most important attributes according to the explanation (either in isolation or in combination) affects the score of the system for a specific prediction. Intuitively, when an explanation is faithful to the ER system, altering the most "salient" attribute in isolation will yield the highest change in the prediction score, altering the second most salient attribute will yield the second highest change in the prediction score, etc., leading to an expected monotonically decreasing change in the prediction score. In Figure 3 the *red line*



Free record

L_Beer (L_B)	L_Factory (L_F)	L_Style (L_S)	L_ABV (L_A)
Manitout Am...	Brewery Terra...	American...	5.30 %

Pivot record

R_Beer	R_Factory	R_Style	R_ABV
Edge Imperial...	Edge Brew...	American...	8.50 %

Score
0.02

Support record

L_Beer	L_Factory	L_Style	L_ABV
Imperial Red	Hop Valley...	American..	8.00 %

Score
0.56

Altered free record — Pivot record

L_Beer	L_Factory	L_Style	L_ABV	R_Beer	R_Factory	R_Style	R_ABV
Imperial Red	Brewery Terra...	American...	5.30 %	Edge Imperial...	Edge Brew...	American...	8.50 %

Score
0.51

Probability of Sufficiency

L_Beer	L_Factory	L_Style	L_ABV
0.08	0.04	0.02	0.03

L_F, L_B	L_S, L_B	L_S, L_F	L_A, L_B	L_A, L_F	L_A, L_S
0.14	0.1	0.03	0.31	0.05	0.01

L_S, L_F, L_B	L_A, L_F, L_B	L_A, L_S, L_B	L_A, L_S, L_F
0.34	0.41	0.33	0.11

Probability of Necessity

L_Beer	L_Factory	L_Style	L_ABV
0.19	0.09	0.04	0.02

Figure 2: Inspecting the effect of altering an attribute in an open triangle for a non-matching prediction by *DeepMatcher*.

shows how the ER system prediction score (y-axis) is affected when altering a single attribute, e.g., altering the most salient attribute *L_Beer* makes the prediction score change from 0.02 to 0.58, altering the second most salient attribute *R_Beer* makes the score change from 0.02 to 0.49 (< 0.58), and so on. When altering the top k salient attributes jointly, we instead expect a monotonically increasing change in the prediction score. We can visualize this as the *green line* in Figure 3: when altering top 2 salient attributes (*L_Beer AND R_Beer*) we obtain a bigger change in the score with respect to when we alter the top 1 salient attribute (*L_Beer*). We name one such visualization as *saliency-effect*.

For counterfactual explanations CERTEM provides common metrics [9] that allow a quantitative explanation evaluation.

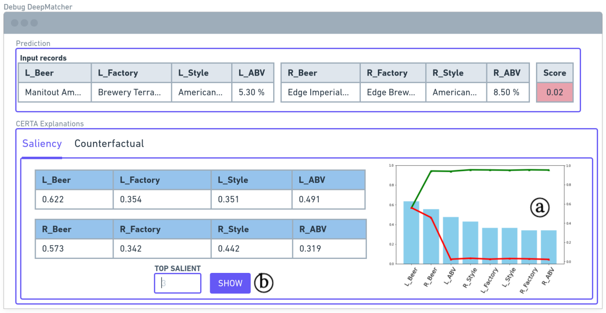
To debug the behavior of an ER system, CERTEM makes a saliency explanation *actionable* as follows: it takes the values corresponding to the top k "salient" attributes and shows the occurrences of such values in positively vs negatively labeled examples in the ground truth. This allows users to more easily detect biases in the training data in terms of imbalanced samples. When no obvious biases or insights arise from debugging the system via saliency explanations, CERTEM makes counterfactual explanations actionable by incorporating them in the original training set and giving users the possibility of retraining the ER system. After retraining the system, CERTEM reports the overall F1 on the test set as well as the new prediction on the same input.

Finally, it is possible in CERTEM to include other explanation systems for head to head comparisons with CERTA outputs.

4 DEMONSTRATION OVERVIEW

We will demonstrate usefulness and effectiveness of CERTEM with respect to three scenarios, using different datasets and ER systems.

For the demonstration, we will use publicly available ER datasets from the DeepMatcher dataset repository.[2] In particular we will consider the following datasets:



Input records

L_Beer	L_Factory	L_Style	L_ABV	R_Beer	R_Factory	R_Style	R_ABV	Score
Manitout Am...	Brewery Terra...	American...	5.30 %	Edge Imperial...	Edge Brew...	American...	8.50 %	0.02

L_Beer	L_Factory	L_Style	L_ABV
0.622	0.354	0.351	0.491

R_Beer	R_Factory	R_Style	R_ABV
0.573	0.342	0.442	0.319

Figure 3: Debugging a *DeepMatcher*'s prediction with saliency explanations.

- *Abt-Buy*: a product dataset from *Abt* and *Buy* online retailers; each record has three attributes.
- *BeerAdvo-RateBeer*: a dataset containing beer data from *BeerAdvocate* and *RateBeer* data sources; each record has four attributes.
- *iTunes-Amazon*: a dataset containing music data from *iTunes* and *Amazon*; each record has eight attributes.

For the sake of clarity we prefix names of attributes coming from U and V with *L_* and *R_* respectively (e.g., *L_Beer* corresponds to the *Beer* attribute from *BeerAdvo* datasource while *R_Beer* corresponds to the *Beer* attribute from *RateBeer* datasource).

Our demonstration will consider predictions made by the following deep learning based ER systems:

- DeepER [5], a system based on the distributed representation of records.
- DeepMatcher [10], a system based on the distributed representation of attributes.
- *Ditto* [7], a system using the Transformer architecture that also adopts data augmentation and injection of domain knowledge.

We will also compare the explanations generated by CERTEM (based on CERTA) to those generated by the following alternative explanation approaches:

- for saliency: Mojito [4], LandMark [2] and SHAP [8].
- for counterfactuals: DiCE [9], LIME-C [11] and SHAP-C [11].

4.1 Demonstration scenarios

***Scenario 1: Context is All You Need*.** We show how to use CERTEM to visualize how different ER systems react to input perturbations. We will allow users to interactively inspect how the perturbations generated via CERTA open triangles affect the prediction when copying values from attributes belonging to the *support record* into corresponding attributes of the *free record* (Figure 2 ⓓ). At the same time we show how CERTA builds lattice structures and accounts for probability of necessity and sufficiency (Figure 2 ⓒ).

Given an open triangle, CERTEM interactively visualizes how perturbing each possible combination of attributes affects the score of the ER system at hand, we seek to highlight what kinds of perturbations have low versus high impact on the score of the model. After having shown a few examples, we will engage the audience

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

and let them guess which attribute perturbations lead to noticeable versus negligible changes in the predicted outcome.

At the end of this part of the demonstration the audience is expected to have gained a more solid "context" about how different ER systems react to changes in the inputs (e.g., *Ditto* is more robust to perturbations than *DeepMatcher*).

Scenario 2: Different Rationales. Building up on the context gained in the previous scenario, we aim at showing how different ER systems can make concordant and correct predictions for different "reasons" and, consequently, they demand different counterfactuals. The goal of this scenario is to emphasize the need for explanations even for cases that look obvious for humans, as explanations might highlight strange or unexpected system behaviors.

We will consider record pairs that are correctly classified (true positives and true negatives) by the ER systems (Figure 1 ⓐ). Leveraging the context gained in the previous scenario, we ask the audience to propose an explanation for each ER system prediction and then show the actual saliency explanation generated by CERTEM. Then, we will compare saliency explanations for the same concordant predictions made by different systems. We will evaluate the following:

- the overlap between saliency explanations for predictions made by different systems, trying to answer the question: *"do different systems make same predictions for different rationales?"*.
- the overlap between the saliency explanations guessed by the audience and the actual saliency explanations generated by CERTA.
- how much the generated explanations are faithful to the model via *saliency-effect* (Figure 3 ⓐ).

We will compare the *saliency-effect* of different explanation systems, demonstrating the superior faithfulness of CERTA saliency explanations of ER systems.

Finally, we will compare the counterfactual explanations generated by CERTEM on the evaluated predictions. We will further investigate how different ER systems can be led to make a wrong prediction for an originally correct prediction by means of counterfactuals. More specifically we will show:

- the overlap between counterfactual explanations for predictions made by different systems;
- some typical performance metrics for counterfactuals like: proximity, diversity and validity [9].

As done for saliency, we will also include other counterfactual generation systems in such an evaluation.

Scenario 3: Counterintuitive Predictions. Another scenario that often demands explanations is when predictions made by ER systems are clearly counterintuitive, with respect to human intuition (e.g., two records are predicted as non-matching, while they are clearly referring to the same entity from a human perspective), or simply wrong with respect to its label in the ground truth.

We will demonstrate how to provide explanations with CERTEM for different record pairs that are wrongly predicted by the considered ER systems while being clearly predictable by humans.

In this scenario we aim at demonstrating how CERTEM can leverage saliency explanations to debug the training data and possibly discover biases that may lead to incorrect predictions.

Additionally, we will demonstrate how CERTEM feeds counterfactual explanations into a counterfactual data augmentation scheme to improve the training data and, consequently, fix wrong classification at instance level. We will start by presenting some wrong/counterintuitive predictions made by the ER systems. We will again engage the audience and ask them to guess what are the most salient attributes with respect to each misclassification. We will then generate both saliency and counterfactual explanations with CERTEM. While a quantitative evaluation on the same metrics presented in *Scenario 2* will be done, in this scenario we put higher emphasis on how CERTEM makes explanations more *actionable.*

Saliency guided debugging. We will use the saliency explanations generated by CERTEM for wrong predictions to navigate through the training set, in search of biases that might have caused such misclassified instances. To do this we will take the top k salient attributes, and show the records that contain the corresponding values (Figure 3 ⓑ). We will look for samples that are highly class imbalanced in the training set, making the question *"do salient attributes' values appear only in negatively (resp. positively) labeled samples in the training set?"*.

Counterfactually augmenting training sets. We will leverage counterfactual explanations generated by CERTEM in order to improve the ER system via a data augmentation scheme. Given a wrongly classified record pair, we will use CERTEM to select a few counterfactual explanations, incorporate them within the training set and retrain the ER system from scratch. We will then check that the overall accuracy of the system hasn't dropped while the previously misclassified input pairs are correctly classified by the ER system.

At this stage, we expect the audience to have gained a better understanding of how CERTEM makes explanations computed by CERTA more actionable in terms of debugging and fixing ER systems, when needed.

REFERENCES

[1] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

[2] A. Baraldi, F. D. Buono, M. Paganelli, and F. Guerra. Using landmarks for explaining entity matching models. In *EDBT*, 2021.

[3] N. Barlaug and J. A. Gulla. Neural networks for entity matching: A survey. *ACM TKDD*, 15(3), 2021.

[4] V. Di Cicco, D. Firmani, N. Koudas, P. Merialdo, and D. Srivastava. Interpreting deep learning models for entity resolution: an experience report using lime. In *aiDM*, 2019.

[5] M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, and N. Tang. Distributed representations of tuples for entity resolution. *PVLDB*, 11(11), 2018.

[6] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[7] Y. Li, J. Li, Y. Suhara, A. Doan, and W. Tan. Deep entity matching with pre-trained language models. *PVLDB*, 14(1):50–60, 2020.

[8] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *NIPS*, 2017.

[9] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *ACM FAT* '20*, 2020.

[10] 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 *ACM SIGMOD*, 2018.

[11] Y. Ramon, D. Martens, F. J. Provost, and T. Evgeniou. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, LIME-C and SHAP-C. *Adv. Data Anal. Classif.*, 14(4), 2020.

[12] T. Teofili, D. Firmani, N. Koudas, P. Merialdo, and D. Srivastava. Effective explanations for entity resolution models. *to appear in ICDE*, 2022.