Reproducibility Report for ACM SIGMOD 2020 Paper: "Towards Interpretable and Learnable Risk Analysis for Entity Resolution"

R. MALINGA PERERA, University of Melbourne, Australia

This reproducibility entry successfully reproduces the main contribution of the paper. Nonetheless, it does not cover the comparative methods reported in the paper nor the supporting extension studies (sensitivity analysis).

1 INTRODUCTION

This reproducibility report assesses the reproducibility of experimental results of [1]. After the environment setup, all the core result generation was performed using a single parameterised command. The experiment successfully reproduced the main results in the paper. Worth pointing out that the environment setup was not clearly explained in the reproducibility scripts and was successfully performed only with the significant help from the authors. The provided script only reproduces the LearnRisk and Baseline results and not the other comparative methods reported in the paper nor the supporting extension studies.

2 SUBMISSION

As a first step of this reproducibility entry, the code has been provided as a downloadable link.[1] After that, as the next step, it requires to run the command below.

$$python\ run.py\ --dataset\ <dataset\_id>\ --split\ <split\_id>\ --epoch\ <\#\_epochs>\ . \quad (1)$$

Between these two steps, we need to install the requirements for this entry. The requirement installation process was not clearly explained in the reproducibility entry. After the environment setup, Command 1 can be executed with different parameters for *dataset* and *split* options to generate the results. This step was smoothly executed without any issue. Possible datasets are 0: DBLP-Scholar, 1: Abt-Buy, 2: Amazon-Google,3: songs, 4: DA2DS, 5: AB2AG and the splits are 0: (1:2:7), 1: (2:2:6), 2: (3:2:5). The third parameter (i.e., the number of epochs) didn't work as expected. We had to perform a code level change to adjust the number of epochs.

3 HARDWARE AND SOFTWARE ENVIRONMENT

The original (paper) environment used a GPU for processing whereas the reproducibility environment used a CPU. The initial experiments in the paper used 1000 epochs, but we had to limit it to 100 epochs due to hardware limitations in the reproducibility environment (the experiments took around 5x longer with CPU processing). The change was performed only after consulting the authors, and we didn't observe any noticeable difference in the results attributed to the reduced number of epochs. This was verified running DBLP-Scholar dataset with 1:2:7 split under 1000 epochs and 100 epochs.

4 REPRODUCIBILITY EVALUATION

4.1 Process

The most challenging part of the process was the environment setup (installing the application dependencies). This process included an installation using a requirements file (requirements.txt) as

[1] https://drive.google.com/file/d/19ehM468OC8cDl3Y0ii45e3KzlBvVN4FJ/view?usp=sharing

Table 1. Hardware & Software environment

	Paper	Repro Review
CPU	Intel i5-9400F	Intel Xeon Platinum 8260
Cores	6	48
GHz	2.9	2.4
RAM	16GB	1.1TB
Storage	HDD	HDD
GPU	GeForce GTX 1660	N/A

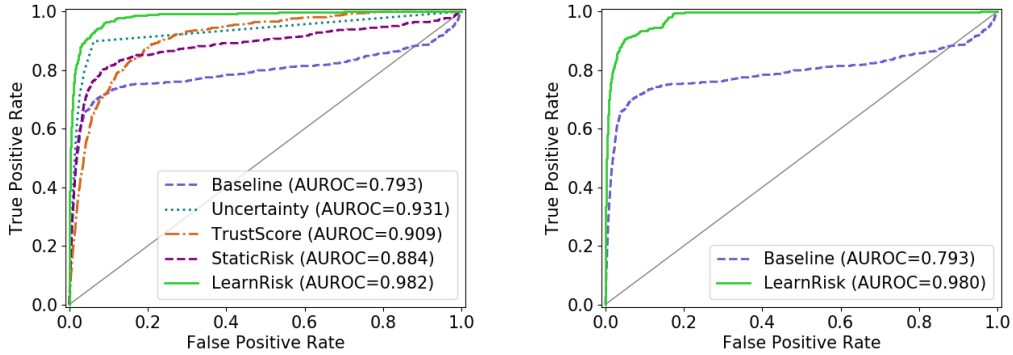
Table 2. Results summary: LearnRisk AUROC value comparison

Dataset	Split	Paper AUROC	Reproduced AUROC
DBLP-Scholar	0	0.982	0.980
	1	0.985	0.984
	2	0.973	0.970
Abt-Buy	0	0.974	0.973
	1	0.954	0.956
	2	0.959	0.963
Amazon-Google	0	0.939	0.939
	1	0.914	0.915
	2	0.930	0.930
Songs	0	0.989	0.989
	1	0.984	0.984
	2	0.992	0.991
DA2DS	2	0.991	0.989
AB2AG	2	0.939	0.936

well as some manual dependency installations. A simple installation using the requirements file was unsuccessful due to missing/incorrect versions provided in the file. After several modifications, the issues were rectified. We recommend the authors to produce a new script, with a clear explanation of the steps the reader should follow. We understand that dependencies make installation somewhat challenging, nonetheless more attention should be paid to this step.

4.2 Results

The reproducibility entry covers the main result of the paper. All LearnRisk AUROC values mentioned in Figure 9 and Figure 10 were successfully reproduced with negligible difference (see Table 2). However, the provided scripts only reproduce the LearnRisk and Baseline results and not the other comparative methods reported in the paper (Uncertainty, TrustScore, StaticRisk) (ex: Figure 1) nor the supporting extension studies (Comparison with HoloClean, Performance Sensitivity, Scalability Evaluation). According to the authors, the comparison methods need different running environments. Consequently, the authors ran them independently and collected their results. The authors did not include this part in the reproducibility entry due to the complexity of the process.



(a) Paper Result **(b) Reproduced Result**

Fig. 1. Paper result vs. reproduced result for DBLP-Scholar dataset with 1:2:7 split

REFERENCES

[1] Zhaoqiang Chen, Qun Chen, Boyi Hou, Zhanhuai Li, and Guoliang Li. 2020. Towards Interpretable and Learnable Risk Analysis for Entity Resolution. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (Portland, OR, USA) *(SIGMOD '20)*. Association for Computing Machinery, New York, NY, USA, 1165–1180. https://doi.org/10.1145/3318464.3380572