Technical Report

Appendix of the Paper: Explaining Entity Matching with Clusters of Words

Riccardo Benassi
DIEF
UNIMORE
Modena, Italy
riccardo.benassi@unimore.it

Francesco Guerra
DIEF
UNIMORE
Modena, Italy
francesco.guerra@unimore.it

Matteo Paganelli
Hasso Plattner Institute
University of Potsdam
Potsdam, Germany
matteo.paganelli@hpi.de

Donato Tiano
DIEF
UNIMORE
Modena, Italy
donato.tiano@unimore.it

This appendix provides additional material on the experimental study in the submitted paper. This extra material has been omitted for space reasons in the submitted version. It consists of the following parts:

- Supporting the use of LIME as internal explainer for CREW (Section I);
- Analyzing the efficiency of CREW when a maximum of 40 words per record pair has been selected (Section II);
- Experimenting with a different implementation of the matching consistency property that relies on BERT embedding similarity instead of word Jaccard similarity (Section III);
- Justifying the properties used for the evaluation of the comprehensibility of the explanations (Section IV).

I. Using CREW coupled with LIME

CREW relies on an external explainer for providing the importance of the words according to a given EM Model as described in line 4 of Algorithm 2 and for weighting the clusters, as introduced in Section III.D. In the paper, we adopted LIME for this purpose. The selection of LIME is motivated by three main reasons:

1) Our previous experience with LIME. In other projects (e.g., WYM [1], Landmark [2]) we used LIME as a baseline, thus giving us some expertize on the project;
2) The explainer is an external component for CREW in charge of generating the impact scores. Any explainer can be used for this purpose, without any changes to the pipeline implemented in the approach.
3) We compared the performance of CREW coupled with LIME against CREW coupled with SHAP, i.e., another well known explainer, used in several approaches. Table I shows the results of the comparison of the degradation scores obtained. We observe that there is no difference between the approaches. CREW coupled with LIME performs slightly better than SHAP in 9 datasets out of 12, and performs slightly less if we weight the clusters with the average technique.

	CREW (SHAP)	CREW (SHAP_AVG)	CREW (LIME)	CREW (LIME_AVG)
D-DA	0.422	0.342	0.247	0.204
D-DG	0.505	0.48	0.563	0.45
D-IA	0.615	0.471	0.656	0.552
D-WA	0.623	0.536	0.624	0.48
S-AG	0.575	0.554	0.643	0.581
S-BR	0.62	0.51375	0.607	0.501
S-DA	0.418	0.376	0.495	0.404
S-DG	0.518	0.503	0.55	0.481
S-FZ	0.449	0.367	0.587	0.514
S-IA	0.595	0.535	0.544	0.529
S-WA	0.64	0.588	0.68	0.537
T-AB	0.666	0.598	0.681	0.541
AVG	0.554	0.489	0.573	0.481

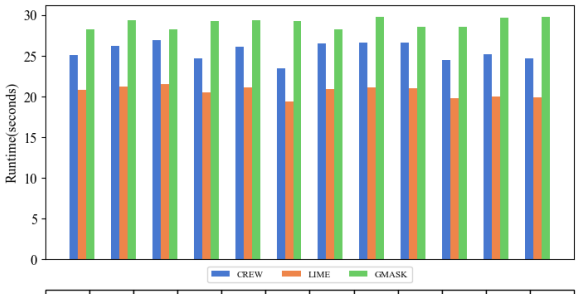
TABLE I: Degradation score: comparison between CREW coupled with SHAP and CREW coupled with LIME

II. Efficiency with a reduced number of words

The time performance evaluation in the paper demonstrates that CREW efficiency is similar to the one of LIME (which has been used in real scenarios) apart from for the S-IA, T-AB and D-IA datasets. The time breakdown shows that the bottleneck is the correlation clustering which takes a long time for these datasets, due to the high number of words. Therefore, we computed again the efficiency, limiting the number of words to 40. This is a number of words that usually allows the generation of useful explanations and is 4 times the number of words that uses the competing approach GMASK. With this limitation, CREW time performance becomes similar in all datasets with an average time equal to 25.53 seconds, as we can see in Figure 1a. Since GMASK takes similar times to compute the explanations we can see that CREW performs approximately 4 times faster than GMASK. The time breakdown in Figure 1b shows again that correlation clustering is the component that is more affected by the amount of words to manage.

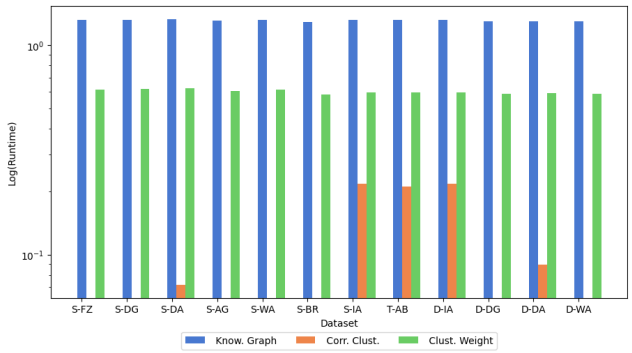
III. Alternative implementation of the matching consistency

In this section we provide an alternative implementation of the matching consistency property, which evaluates the quality



	S-FZ	S-DG	S-DA	S-AG	S-WA	S-BR	S-IA	T-AB	D-IA	D-DA	D-DG	D-WA
#Words	24.88	31.89	35.76	19.53	33.47	26.27	39.91	39.66	24.01	36.31	32.25	33.47

(a) Average time (in seconds) required to generate an explanation.



(b) Time breakdown along knowledge graph creation, correlation clustering computation and clustering weighting.

Fig. 1: CREW efficiency by limiting the maximum number of words per record pair to 40 words.

of an explanation for a matching task. Intuitively, this property measures the ability of a cluster-based explanation to group words that refer to similar concepts within the same cluster and divide dissimilar concepts into different clusters. The original implementation of this property uses the Jaccard similarity to evaluate the similarity between words: words with a similarity greater than 0.9 are considered to refer to similar concepts, while words with similarity less than 0.1 are considered to refer to dissimilar concepts.

In this section we replace the Jaccard similarity with the cosine similarity between BERT embeddings. Since CREW uses this information internally to group semantically related words and increase the comprehensibility of the explanation for the user, this evaluation is not completely fair however it can provide interesting insights into the real contribution of semantic relatedness between words in CREW's final outcome. Remember that the explanations produced by CREW are the result of three signals that are processed jointly and whose isolated contribution is difficult to estimate. This experimentation is therefore also intended to evaluate this aspect.

The results of this evaluation are reported in Table II, where we compare the matching consistency scores obtained by CREW with the competitor approaches (i.e. LIME, Mojito and GMASK) and in the ablation study. As can be seen, there is no significant variation in the results when compared to the Jaccard-based implementation. The new implementation of matching consistency generates the following average results for the various approaches analyzed, i.e., CREW: 0.716, LIME: 0.52, Mojito: 0.621, GMASK: 0.514, while the Jaccard-based implementation the following ones: CREW: 0.713, LIME : 0.528, Mojito: 0.623, GMASK: 0.512. This is presumably motivated by the fact that the selection of words with cosine similarity greater than 0.9 (and dually less than 0.1) already corresponded with the selection of more (less) syntactically similar words.

IV. Evaluating the comprehensibility of the explanations

In Section IV.A of the paper, we introduce four properties to evaluate the comprehensibility of the explanations. The properties emerged from some evidences we observed thanks to our experience in the field. In particular:

1) **Matching consistency.** We see that an explanation of a matching decision is understandable when it is consistent with two principles: a) it labels similar text portions in the pair of entity descriptions as evidence of match, and, conversely, b) it labels dissimilar text portions in the two entity descriptions as evidence of non-match.
2) **Semantic cohesion.** We noticed that words belonging to the same attribute of an entity description and the corresponding matching attribute from the second description have to exhibit high semantic cohesion, as they describe the identical property of the entity.
3) **Intra-cluster importance cohesion.** An explanation composed of groups of words is comprehensible if the words within the same group are semantically related and have the same importance.
4) **Inter-cluster importance cohesion.** An explanation composed of groups of words cohesion could be complex to understand if all groups have the same importance. In this case, users cannot discriminate the most important elements from the descriptions. Clusters should then show a different importance.

To support the claim that the aforementioned properties are important for evaluating the comprehensibility of the explanations, we delivered a questionnaire (see Figure 2) to 18 experts from our organizations.

The results of our small scale evaluation demonstrate show that the experts

1) prefer group-level (66.67%) and attribute-level (27.78%) explanations;

TABLE II: Evaluation of the matching consistency property of the explanations generated by CREW compared with competitor approaches (left) and in the ablation study (right).

Dataset	MC↑				MC (Ablation)↑			
	CREW	LIME	Mojito	GMASK	CREW	w/o Sim.	w/o Sch.	w/o Imp.
S-FZ	**0.798**	0.535	0.693	0.520	0.798	0.697	0.663	0.921
S-DG	**0.695**	0.530	0.661	0.510	0.695	0.657	0.617	0.798
S-DA	**0.752**	0.526	0.686	0.511	0.752	0.684	0.647	0.816
S-AG	**0.653**	0.537	0.601	0.516	0.653	0.606	0.608	0.768
S-WA	**0.703**	0.511	0.625	0.514	0.703	0.631	0.628	0.789
S-BR	**0.768**	0.555	0.625	0.538	0.768	0.659	0.724	0.856
S-IA	**0.894**	0.529	0.711	0.517	0.894	0.748	0.823	0.946
T-AB	**0.659**	0.505	0.596	0.503	0.659	0.608	0.597	0.774
D-IA	**0.664**	0.518	0.537	0.504	0.664	0.590	0.671	0.783
D-DA	**0.716**	0.524	0.584	0.514	0.716	0.594	0.697	0.729
D-DG	**0.632**	0.531	0.551	0.513	0.632	0.592	0.612	0.690
D-WA	**0.653**	0.520	0.578	0.508	0.653	0.599	0.601	0.723
AVG	0.716	0.527	0.621	0.514	0.716	0.639	0.657	0.799
Δ%	-	-	-	-	-	**+10.75**	+8.24	-11.59

2) agree that matching entities should share common sequence of words (88.89%);
3) completely agree that descriptions of non-matching entities cannot contain common sequence of words (100%);
4) are aware that the dataset structure is important for supporting the decision (77.17%);
5) prefer explanations showing the most important element contributing to the decision only (72.22%)
6) prefer a large number of small groups of words (88.89%).

This analysis is too small to be considered more than a slight indication. However it confirms what we learn from our experience and what we think it is the common sense in the definition of a comprehensible explanation.

REFERENCES

[1] Andrea Baraldi, Francesco Del Buono, Francesco Guerra, Matteo Paganelli, and Maurizio Vincini. An intrinsically interpretable entity matching system. In *EDBT*, pages 645–657. OpenProceedings.org, 2023.

[2] Andrea Baraldi, Francesco Del Buono, Matteo Paganelli, and Francesco Guerra. Using Landmarks for Explaining Entity Matching Models. In *EDBT*, 2021.

Song	Artist	Genre	Time	Release
Stars Come Out Mason Remix Dance Electronic	/	/	5:49	20-May-14

Song	Artist	Genre	Time	Release
Stars Come Out Francis Remix	/	Dance & Electronic	4:08	May 20 2014

Entity Matching aims to determine if two entries in a dataset refer to the same real-world entity. For example, the pair of records shown in the next Figure represent the description of two songs extracted from the Dirty iTunes-Amazon dataset.
The Entity Matching problem applied to the records in the Figure aims to decide if the entries describe (or don't describe) the same song.

Real-world scenarios typically require that an explanation be provided as to why the Entity Matcher decided to classify the entries as matching (or not matching). The explanation is usually a score assigned to the entity descriptions (or part of them) that shows their importance in the decision.

Question 1)
Which is the level of granularity that an explanation should have?

a) word level (each word from each entity description should have an importance score)
b) group of words (words from both the entity descriptions have to be grouped according to some criterium and then the score is at the group level)
c) dataset attribute (each attribute in the dataset should have an importance score)
d) other

Question 2)
Do you think that descriptions of matching entities should share common sequence of words? E.g., the titles of two song descriptions have to contain the sequence of words "Stars come out".

Question 3)
Do you think that "in general" descriptions of non-matching entities cannot contain common sequence of words?

Question 4)
Do you think that the structure of the dataset in attributes can provide some information for supporting the decision? For example, in matching descriptions the attribute Song Title should contain the same value (or a similar value) for both the entries. Conversely, in non-matching descriptions the attributes should contain different values?

Question 5)
Is it more useful an explanation where all the terms have a close impact score or where only a few terms have a high score (i.e., the explanation highlights the most important elements used for taking the decision)?

Question 6)
Let us suppose that an explanation system provides explanations for groups of words. Do you prefer that the system generates a few large groups of words (even if the words in the same groups can have different meanings and not to be strictly semantically related) or a large number of small groups of semantically related words?

Fig. 2: Questionnaire delivered to experts for supporting the choice of the properties.