Enhanced Benchmark Datasets for a Comprehensive Evaluation of Process Model Matching Techniques

Muhammad Ali[1,2] and Khurram Shahzad[1,2(✉)]

[1] Software Development and Maintenance Center, University of Gujrat, Gujrat, Pakistan
muhammad.ali@uog.edu.pk, khurram@pucit.edu.pk
[2] Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan

Abstract. Process Model Matching (PMM) refers to the automatic identification of corresponding activities between a pair of process models. Recognizing the pivotal role of PMM in numerous application areas a plethora of matching techniques have been developed. To evaluate the effectiveness of these techniques, researchers typically use PMMC'15 datasets and three well-established performance measures, precision, recall and F_1 score. The performance scores of these measures are useful for a surface level evaluation of a matching technique. However, these overall scores do not provide essential insights about the capabilities of a matching technique. To that end, we enhance the PMMC'15 datasets by classifying corresponding pairs into three types and compute performance scores of each type, separately. We contend that the performance scores for each type of corresponding pairs, together with the surface level performance scores, provide valuable insights about the capabilities of a matching technique. As a second contribution, we use the enhanced datasets for a comprehensive evaluation of three prominent semantic similarity measures. Thirdly, we use the enhanced datasets for a comprehensive evaluation of the results of twelve matching systems from the PMM Contest 2015. From the results, we conclude that there is a need for developing the next generation of matching techniques that are equally effective for the three types of pairs.

Keywords: Business process management · Process Model Matching PMMC'15 datasets · Enhanced datasets · Comprehensive evaluation

1 Introduction

Process Model Matching (PMM) refers to the automatic identification of activities between a pair of process models that exhibit the same or similar behavior [1, 2]. The participating activities are called corresponding activities and the pair is called corresponding pair [2, 3]. The identification of corresponding activities has a pivotal role in various applications domains, such as process querying, clone detection, and harmonization of process models [4–6]. Recognizing that, a plethora of PMM techniques have been developed [7].

R. Pergl et al. (Eds.): EOMAS 2018, LNBIP 332, pp. 107–122, 2018.
https://doi.org/10.1007/978-3-030-00787-4_8

To evaluate each of these matching technique, leading experts of the BPM domain have developed three benchmark datasets, formally called PMMC'15 datasets [8]. Since 2015, these datasets are widely used for the evaluation of PMM techniques [9], by using three well-established performance measures, precision, recall and F_1 score [7–9]. The performance scores of these measures are useful for a surface level evaluation of a matching technique. However, our synthesis of the PMMC'15 datasets and the evaluation results have revealed two interrelated issues regarding the evaluation of matching techniques. Prior to discussing the issues, in the remaining part of this section, we first highlight the diversity that can possibly exist in the corresponding pairs. Subsequently, in Sect. 1.2 we discuss the two issues that arise during the evaluation of a matching technique. Finally, in Sect. 1.3 we present the conceptual bases, from text process literature, that we have used for classifying the corresponding pairs.

1.1 Illustration of Diversity in Corresponding Pairs

Figure 1 illustrates the possible diversity between corresponding pairs using admission process models of two universities, University A and B. The diversity represents the varying levels of differences in the formulation of participating labels. In the figure, the corresponding pairs of the two process models are highlighted with grey shades. Note, we have used three different shades of gray color, light gray, ordinary gray and dark gray, to represent the diversity in corresponding pairs. The higher the difference in formulation of labels the higher is the darkness of the color.

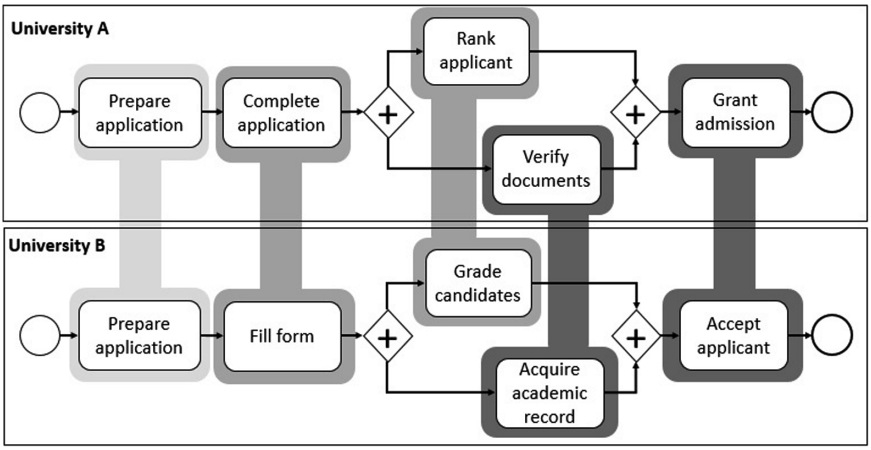

Fig. 1. Illustration of process model matching

In the example, there is no difference in the formulation of label 'prepare application' in the two process models. Due to the absence of this difference, this correspondence is highlighted with light gray color. Similarly, the two labels 'complete application' and 'fill form' are formulated quite differently, i.e. the words are replaced with their synonyms. Due to the slight difference in formulation of labels this

correspondence is highlighted with ordinary gray color. Finally, the formulation of the two labels 'grant admission' and 'accept applicant' is completely different but their business semantics are the same. Due to this significant different in formulation of labels this correspondence is highlighted with dark gray.

The example illustrates the varying differences that may exist in corresponding pairs. In the presence of this diversity, an ideal matching technique should achieve a surface level performance scores as well as comparable performance scores for the corresponding pairs of all three shades.

1.2 Motivation for Enhancing the Benchmark Datasets

Leading experts from the BPM domain introduced three real-world datasets for the evaluation of process model matching techniques, formally called PMMC'15 datasets [1, 9]. The three datasets are named as, University Admissions (UA), Birth Registration (BR) and Asset Management (AM) datasets [1, 8]. Each dataset is composed of a collection of process models and gold standard correspondences between pairs of process models. The UA, BR and AM datasets are composed of 9, 9 and 72 real-world process models, respectively. Each dataset has 36 process model pairs and gold standard correspondences between activities of the 36 pairs. The detailed specification of the three datasets is presented in Table 1. We consider it important to clarify that these numbers are generated without any pre-processing on the datasets, and that our enhancements to the datasets will change these numbers. From the table it can be observed that UA dataset has 1575 pairs, BR dataset has 633 pairs and AM dataset has 799 pairs.

Table 1. Specification of the PMMC'15 datasets

	UA	BR	AM
Total no of pairs in the dataset	1575	633	799
Number of corresponding pairs	202	183	151
Number of trivial corresponding pairs	136	70	102
Number of non-trivial corresponding pairs	66	113	49

The PMMC'15 datasets have been used in numerous studies for the evaluation and comparison of process model matching techniques [1–3, 8, 10]. All these studies rely on a single Precision, Recall and F_1 score to represent the effectiveness of a technique. Consequently, a matching technique with higher F_1 score is declared as more effective than the one with lower F_1 score. However, there are two interrelated issues with this combination of datasets and performance measures. In the presence of these issues a more thorough evaluation of the matching techniques is desired, before pronouncing a matching technique more effective than the other one. These issues are as follows:

– *Inflated F_1 score:* Our syntheses of the PMMC'15 datasets have revealed that a significant percentage of corresponding pairs in all the three datasets are either identical or similar. We formally refer to these pairs are *'trivial'* corresponding

pairs. It is because, in these pairs, either there is no difference in the formulation of the participating labels or the change is as small as changing the form of a word and/or adding a stop word. From the table it can be observed that out of the 202 corresponding pairs in the UA dataset, 136 are *trivial* corresponding pairs. Similarly, in BR dataset 70 out of 183 pairs, and in AM dataset 102 out of 151 pairs are *trivial* corresponding pairs. Therefore, the inclusion of such a large percentage of *trivial* corresponding pairs artificially inflate the F_1 score achieved by a matching technique.

– *Surface level Evaluation:* As illustrated in the preceding subsection, the PMMC'15 datasets contain diverse corresponding pairs. In the presence of this diversity, the use of a single value of each performance measure, for the complete dataset, provides a valuable surface level evaluation of a technique. However, a single score does not provide important insights about the behavior of a matching technique, which essentially requires answers questions like, how effective is the technique for the diverse corresponding pairs, Light Gray, Ordinary Gray and Dark Gray pairs.

Based on the above discussion we conclude that there is a dire need to enhance the PMMC'15 datasets by classifying corresponding pairs based on the diversity of the pairs. Furthermore, in addition to the surface level performance scores, due attention should be paid to the performance scores of the diverse corresponding pairs.

1.3 Conceptual Bases for Classifying Corresponding Pairs

Several studies in the natural language processing domain have identified three language independent relationships between a text pair, depending upon the level of similarity between the two texts in the pair [11–13]. These relationships have been widely used for several text processing tasks, such as, plagiarism detection [12], text reuse in journalism [11, 13] and duplicate document identification [14]. The relationships are Near Copy, Light Revision, and Heavy Revision. A brief overview of each type of relationship is as follows:

– *Near Copy:* The two texts are called Near Copy of each other if one text can be generated by slightly rephrasing the other text. That is, by adding stop words or changing the form of the word. A possible near copy of 'best student' is 'the best student'.
– *Light Revision:* The two texts are called Light Revision of each other if one text can be generated by substantially paraphrasing the other text. That is, by replacing words with synonyms, or adding additional words, etc. A possible Light Revision of 'best student' is 'outstanding undergrad student'.
– *Heavy Revision:* The two texts are called Heavy Revision of each other if one text can be generated by significantly paraphrasing the other text. That is, by replacing words with alternate words, reordering the words or making any other change in which the semantic meanings of the text are not changed. A possible Heavy Revision of 'best student' is 'topper of the class'.

The rest of the paper is organized as follows: in Sect. 2 we present the details of the changes we have made to enhance the PMMC'15 datasets. In Sect. 3 we present the

enhanced dataset for the evaluation of three semantic matching measures. In Sect. 4 we present the use of the enhanced dataset for the evaluation of 12 matching systems from the PMM Contest 2015. Finally, Sect. 5 concludes the paper.

2 Enhancing the PMMC'15 Datasets

This section presents the first contribution of our work, enhancing the PMMC'15 dataset for a comprehensive evaluation of PMM techniques. To that end, in this section, we first introduce the three types of pairs that are used to represent the diversity in corresponding pairs. Secondly, we discuss the preprocessing that we have performed on the datasets. Finally, we present the procedure that we have used for enhancing the datasets.

2.1 Representing Diversity in Corresponding Pairs

We propose three types of pairs to represent diversity in corresponding pairs. These types stem from the three types of relationships between text pairs, presented in Sect. 1.3, and the synthesis of the PMMC'15 datasets. The three types that we have used for classifying corresponding pairs are, Verbatim, Modified Copy and Heavy Revision. A brief overview of each type is as follows:

– *Verbatim (VB):* A corresponding pair is classified as Verbatim if the two labels in the pair are similar or almost the same. Based on the definition of Near Copy relation in a text pair as well as the synthesis of the PMMC'15 datasets, we have identified three criteria to declare a pair Verbatim. The three criteria and their examples from the PMMC'15 are presented in Table 2.

Table 2. Criteria and examples of declaring a pair Verbatim

Criteria	Examples
Identical label	Creation birth certificate - create birth certificate
Identical label without stop words	Receive notification birth - receive notification of birth
Reordered words Identical, but	Check nationality of parents - check parent's nationality

– *Modified Copy (MC):* A corresponding pair is classified as Modified Copy if the two labels in the pair have the same semantic meanings but the formulation of the labels is substantially different. Based on the definition of Light Revision relation in a text pair as well as the synthesis of the PMMC'15 datasets, we have identified three criteria for declaring a pair Modified Copy. The three criteria and their examples from the PMMC'15 are presented in Table 3.

Table 3. Criteria and examples of declaring a pair Modified Copy

Criteria	Examples
Adding/deleting a few words	Register child - register baby as German 1
Replacing synonyms	Confirm identity - check identity
Switching labeling style	Register child - child registration

– *Heavy Revision (HR):* A corresponding is classified as Heavy Revision if the formulation of the two labels is significantly different, or one label subsumes the other label. We have identified two criteria for declaring a pair Heavy Revision. The criteria and their examples from the PMMC'15 are presented in Table 4.

Table 4. Criteria and examples of declaring a pair Heavy Revision

Criteria	Examples
Substantially revised labels	Receive documents - receive the citizen decision
Subsume the other label	Register child - child registration

2.2 Pre-processing the PMMC'15 Datasets

Prior to annotating a type to a corresponding pair, we also synthesized the publicly available[1] results of 12 matching systems as well as the gold standard[2] included in the results. The synthesis revealed two types of discrepancies in the gold standard that must be omitted before annotating a type to each corresponding pair. These discrepancies are as follows:

– There are 188 corresponding pairs in the gold standard of the UA dataset that do not have a meaningful label. For example, 'IntermediateCatchEvent' – 'IntermediateCatchEvent', and 'ExclusiveGateway' – 'ExclusiveGateway'.
– In each of the three datasets, there are several corresponding pairs that have the same business impact, but they are declared as unequivalent in the gold standard. Examples of these pairs are as follows: 'wait for results' – 'wait for results' and 'clearing is posted' – 'clearing is posted'. The amount of these pairs in the UA, BR and AM datasets are 13, 42 and 213, respectively.

In the first step of the pre-processing, the unlabeled pairs in the UA dataset were removed. In the second step of the pre-processing, discrepancies among the equivalent pairs were rectified in the three datasets. That is, 13 activity pairs for UA dataset, 42 activity pairs for BR dataset and 213 activity pairs for AM dataset were corrected. Accordingly, the UA, BR and AM datasets that we used for annotation was composed of 360, 423 and 456 corresponding pairs, respectively.

2.3 Annotating Types to Corresponding Pairs

We have annotated a type to each corresponding pair in the pre-processed dataset. The three types that we have used for the annotations are, VB, MC and HR. For the annotations we rely on the classification criteria presented in Tables 2, 3 and 4.

As a first step, each corresponding pair was independently annotated by two researchers using the classification criteria. Secondly, the annotations were compared and conflicts were identified. Subsequently, all the conflicts were resolved by a

[1] The results can be downloaded from https://ai.wu.ac.at/emisa2015/contest.php.

[2] Gold standard refers to the benchmark correspondences generated by BPM experts.

consensus approach, that is, by individually discussing each conflicting pair. As an outcome of this activity, all the corresponding pairs were annotated with a mutually agreed pair type, VB, MC or HR. Table 5 shows the distribution of pairs according to types. From the table it can be observed that a significant number of pairs are annotated as VB or HR. However, there are fewer pairs that are annotated as MC. This imbalance in the number of pairs in the three types, reinforces that a single Precision, Recall or F_1 score is not sufficient for a fair evaluation of the PMM technique. Hence, in the rest of the paper, we separately compute the performance scores for individual pair types, in addition to the overall performance scores.

Table 5. Distribution of corresponding pairs according to types

Datasets	VB pairs	MC Pairs	HR pairs	Total
UA	106	53	201	360
BR	125	79	219	423
AM	322	25	109	456
Total	553	157	529	1,239

3 Evaluation of Semantic Similarity Measures

This section presents our second contribution, a comprehensive evaluation of three prominent the semantic similarity measures, using the enhanced PMMC'15 dataset. Below, we first introduce the three semantic similarity measures. Subsequently, we present an overview of the experimental setup and analysis of the results.

3.1 Semantic Similarity Measures

WordNet is a well-established lexical database for English language that is widely used to computing semantic similarity between two concepts [15]. The database consists of over 150,000 nouns, verbs, adverbs and adjectives. The concepts are organized into related synonyms, also called synsets [16, 17]. In addition to the synsets, the concepts in WordNet are linked with each other via a variety of relationships, such as is-a and part-of relationships, to form a network of concepts.

For this study, we have selected three prominent semantic similarity measures that are previously in PMM literature. These measures are, Lin [18], similarity [19] and Path similarity [20]. A brief overview of each similarity measure is as follows:

Lin Similarity. This similarity measure computes similarity between concepts based on the Information Content (IC) of Least Common Subsumer (LCS) in the WordNet database [17]. Subsequently, the similarity of a label pair is calculated by averaging of all optimal words pairs. Formally, word level Lin score is computed by using Eq. 1.

$$Lin(c1, c2) = \frac{2 * IC(LCS(c1, c2))}{IC(c1) + IC(c2)} \quad [16] \tag{1}$$

Lesk Similarity. This similarity measures computes the degree of similarity between two words by calculating the overlap in the dictionary definition of the two words [19, 21]. Subsequently, the similarity of a label pair is calculated by averaging of all optimal words pairs' Lesk value.

Path Similarity. This similarity measure uses the shortest path between two words in the WordNet database to compute similarity between two labels, by using Eq. 2 [20].

$$Sim(L1, L2) = \frac{\sum_{w1 \in L1 \setminus L2} \max(\partial(w1, w2) | w2 \in L2 \setminus L1)}{|L1 \setminus L2|} \quad (2)$$

Where $\partial(w1, w2)$ is path similarity value of words pair w1 and w2 from WordNet.

3.2 Experimentation and Analysis of Results

We implemented the three semantic similarity measures in Python and used them for the experimentation. Each implemented similarity measure takes input a set of activity pairs and returns similarity scores of each input pair. Experiments are performed using the complete dataset (including all pairs in the dataset), as well as using the three types of pairs, separately. The results of the complete datasets provide a surface level evaluation of the matching technique whereas, the results of each type of pair provide valuable insights about the capabilities of the matching techniques. Similarly, separate experiments are performed for each dataset, UA, BR and AM dataset.

The semantic measures return a similarity score between 0 and 1, whereas the performance measures, precision, recall and F_1 score, requires binary decisions, 'Yes' and 'No'. For a technique β, the decision 'Yes' represents that the technique β has declared the pair as corresponding pair (equivalent pair), whereas the decision 'No' represents that the technique has declared the pair as unequivalent pair. To convert the similarity scores between 0 and 1 to Yes and No, we have used a cut-off threshold 0.75. The choice of cut-off threshold stems from the fact that multiple matching systems participated in latest episode Process Model Matching Contest 2015 have shown promising results at this threshold or a similar threshold [8]. The overall performance scores and the performance of each individual types of pairs are presented in Table 6.

Note, for the complete dataset, we have presented all the performance scores in the table. In contrast, for the three types of pairs, we have only presented the Recall scores because the precision scores of all techniques for all types of pairs are 1, due to the absence of unequivalent pairs. A further analysis of the results are as follows:

Overall Results of the Techniques. From the overall results presented in Table 6, it can be observed that there is no significant difference between the performance of techniques for the UA dataset. That is, the F_1 score achieved by the three techniques for UA dataset are comparable. The similarity trend can be observed for the other two datasets. However, the performance scores achieved by all the techniques for AM dataset are higher than BR dataset. Furthermore, the performance scores achieved by all the techniques for BR dataset are higher than UA dataset, indicating that BR dataset contains harder-to-detect pairs than UA dataset. From the table it can also be observed that the cause of below-par F_1 scores is the lower Recall scores. These lower Recall

Table 6. Results of the semantic similarity measures.

Datasets	Measures	Overall			VB	MC	HR
		P	R	F_1	R	R	R
UA dataset	Lin	0.86	0.55	0.67	1	0.66	0.28
	Lesk	0.85	0.55	0.66	1	0.66	0.28
	Path	0.90	0.49	0.63	0.99	0.17	0.30
BR dataset	Lin	1	0.35	0.52	0.96	0.33	0.01
	Lesk	1	0.34	0.50	0.96	0.27	0.01
	Path	0.98	0.30	0.47	0.88	0.17	0.03
AM dataset	Lin	0.93	0.77	0.84	1	0.36	0.18
	Lesk	0.94	0.76	0.84	1	0.36	0.14
	Path	0.99	0.73	0.84	0.98	0.32	0.11

scores represent that there is a need for considering other similarity measures for accurate identification of corresponding pairs.

Performance Variation Across Pairs. The three graphs presented in Figs. 2 and 3 show a performance comparison of the techniques across the three types of pairs. From the figure it can be observed that the Recall for VB pairs is either exactly 1 or nearly 1. This indicates that all the three techniques successfully detected the VB pairs with a very high accuracy. It can also be observed from the graphs that the Recall drops significantly for MC pairs and it becomes extremely low for the HR pairs. This indicates that the similarity measures only identified a fraction of the corresponding pairs in which the constituent labels are substantially different. However, these measures completely failed in identifying the HR pairs. These dropping scores further represent that the enhancements to our dataset are in-line with our plan.

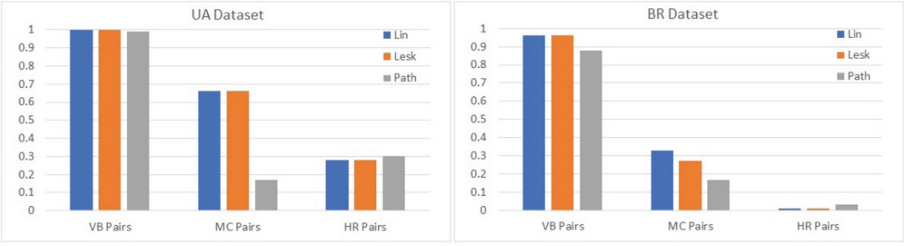

Fig. 2. Performance variation across pairs for UA and BR datasets

Performance Variation Across Techniques. To understand the performance variation across techniques, Fig. 4 plots the average of the Recall scores of the three datasets. From the graph it can be observed that there is no significant difference between performances of the three techniques for all the three types of pairs. This indicates there is no universally acceptable similarity measure that performs equally well for all the three datasets.

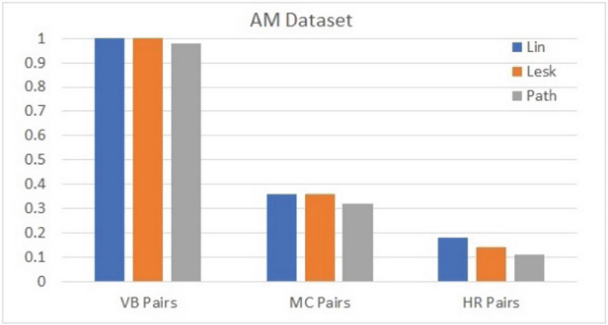

Fig. 3. Performance variation across pairs for AM dataset

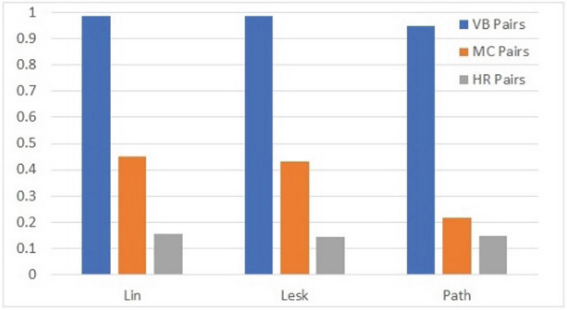

Fig. 4. Performance variation across techniques

Based on the results we conclude the following:

- The MC and HR pairs are composed of hard-to-detect corresponding pairs and the three semantic similarity measures do not show any promise to identify these pairs. We therefore conclude that there is a need for a next generation of matching techniques that can show promising results for MR and HR pairs.
- There is no universal similarity measure that show promising results for all the three datasets.

4 Evaluation of Matching Systems from PMMC 2015

This section presents our third contribution, a comprehensive evaluation of 12 matching systems that participated in Process Model Contest 2015. Similar to the evaluation of semantic similarity measures, we have used our enhanced PMMC'15 dataset for the evaluation of the matching systems. To that end, we mapped the publicly available results[3] to our enhanced dataset, and used it to generate the scores of the performance

[3] The results can be downloaded from https://ai.wu.ac.at/emisa2015/contest.php.

Table 7. Results of UA dataset

	Overall			VB	MC	HR
	P	R	F1	R	R	R
AML-PM	0.56	0.70	0.62	0.93	0.47	0.64
BPLang	0.57	0.44	0.49	0.80	0.21	0.30
NHCM	0.88	0.44	0.58	0.92	0.28	0.22
NLM	1.00	0.29	0.45	0.99	0.00	0.00
MSSS	0.90	0.31	0.46	0.95	0.17	0.00
OP-BOT	0.76	0.37	0.49	0.71	0.28	0.21
KMS	0.77	0.52	0.62	0.99	0.70	0.23
SMSL	0.65	0.38	0.48	0.82	0.19	0.20
TripleS	0.56	0.34	0.42	0.99	0.13	0.05
Knoma – Proc	0.70	0.62	0.66	0.99	0.30	0.51
VM2	0.41	0.58	0.48	0.86	0.51	0.44
pPALMDS	0.32	0.73	0.45	0.96	0.57	0.64

measures. For a thorough evaluation, we generated the scores of the performance measures using the complete datasets as well as for each type of pair, separately.

The results of all techniques for the three datasets are present in Tables 7, 8 and 9. Similar to the results of the semantic measures for the complete dataset, we have presented all the performance scores in the table. In contrast to that, for the three types of pairs we have only presented the Recall scores. Below, we present the *analysis of the results:*

Overall Results of the Techniques. For the UA dataset, overall highest F_1 score of 0.66 is achieved by Knoma-Proc, whereas pPALMDS achieved an F_1 score of 0.45. For the BR dataset, overall highest F_1 score of 0.68 is achieved by pPALMDS, whereas

Table 8. Results of BR dataset

	Overall			VB	MC	HR
	P	R	F1	R	R	R
AML-PM	0.82	0.45	0.58	0.91	0.44	0.18
BPLang	0.94	0.35	0.51	0.77	0.34	0.11
NHCM	0.97	0.36	0.53	0.75	0.47	0.10
NLM	1.00	0.24	0.39	0.77	0.06	0.00
MSSS	1.00	0.22	0.36	0.68	0.09	0.00
OP-BOT	0.92	0.44	0.60	0.70	0.38	0.32
KMS	0.97	0.28	0.43	0.68	0.25	0.05
SMSL	0.72	0.37	0.49	0.78	0.37	0.15
TripleS	0.92	0.35	0.51	0.79	0.39	0.08
Knoma – Proc	0.86	0.37	0.52	0.78	0.46	0.11
VM2	0.69	0.59	0.64	0.78	0.63	0.47
pPALMDS	0.85	0.57	0.68	0.77	0.58	0.46

Knoma-Proc achieved an F_1 score of 0.52. For the AM dataset, overall highest F_1 score of 0.82 is achieved by Knoma-Proc, whereas pPALMDS achieved an F_1 score of 0.44. These results indicate there is no universal system that achieved higher accuracy for all the three datasets.

From the results of the UA dataset it can be observed that KMS show promising results for MC pairs (R = 0.70). However, this technique performs poorly for HR pairs (R = 0.23). From the results of the BR dataset it can be observed that VM2 shows promising result for MC pairs (R = 0.63), and its performance reduces slightly for HR pairs (R = 0.47). Similar trends can be observed for the AM dataset. Based on these results we conclude that, a large majority of the techniques do not show comparable performance for MC and HR pairs.

Table 9. Results of AM dataset

	Overall			VB	MC	HR
	P	R	F1	R	R	R
AML-PM	0.86	0.31	0.46	0.34	0.56	0.17
BPLang	0.79	0.29	0.42	0.32	0.40	0.16
NHCM	0.97	0.25	0.40	0.33	0.20	0.04
NLM	1.00	0.24	0.39	0.34	0.00	0.00
MSSS	0.91	0.23	0.37	0.33	0.04	0.00
OP-BOT	0.68	0.31	0.42	0.34	0.36	0.19
KMS	0.77	0.31	0.44	0.34	0.40	0.19
SMSL	0.79	0.13	0.22	0.15	0.12	0.06
TripleS	0.73	0.32	0.44	0.34	0.32	0.24
Knoma – Proc	0.85	0.79	0.82	0.97	0.60	0.28
VM2	0.74	0.29	0.42	0.31	0.40	0.20
pPALMDS	0.52	0.38	0.44	0.34	0.56	0.47

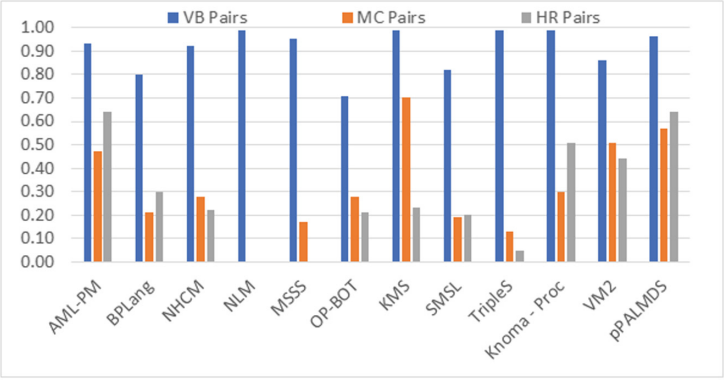

Fig. 5. Performance variation across pairs for UA dataset

Below, we further analyze the results of the matching systems.

Performance Variation Across Pairs. The three graphs in Figs. 5, 6 and 7 plots the Recall score across the 12 matching systems. From the figure it can be observed that the Recall scores for VB pairs are very high for UA and BR datasets. It can also be observed that the Recall scores drop significantly for the MC pairs. Furthermore, the Recall scores drop further for the HR pairs. These results indicate that majority of the matching systems fail to identify the HR pairs. However, there some exceptions (AML-PM, BPLang, Knoma-Proc, and pPALMDS) that achieve higher Recall score for HR pairs than MC pairs for one dataset, UA dataset. Among these, pPALMDS is the extraordinary matching system due to three reasons, (a) for the UA dataset, the matching system achieved higher Recall for the HR pairs than for the MC pairs (0.64 > 0.57), (b) for the BR dataset, the performance decline from MC to HR pairs is not substantial, i.e. 0.12, and (c) for the AM dataset, the performance decline from MC to HR pairs is not substantial, i.e. 0.09. Hence, we declare pPALMDS as the best performing system.

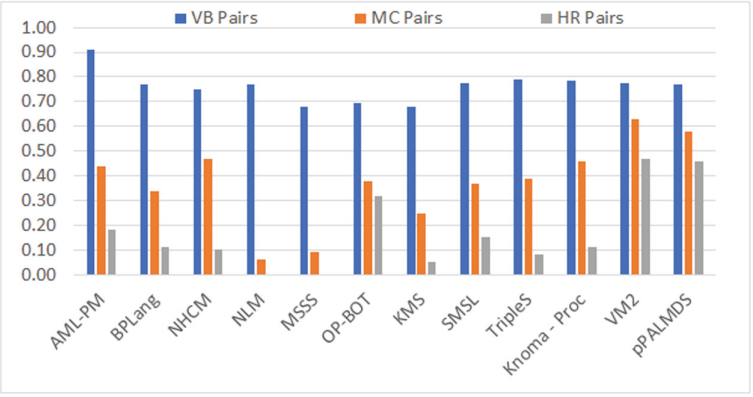

Fig. 6. Performance variation across pairs for BR dataset

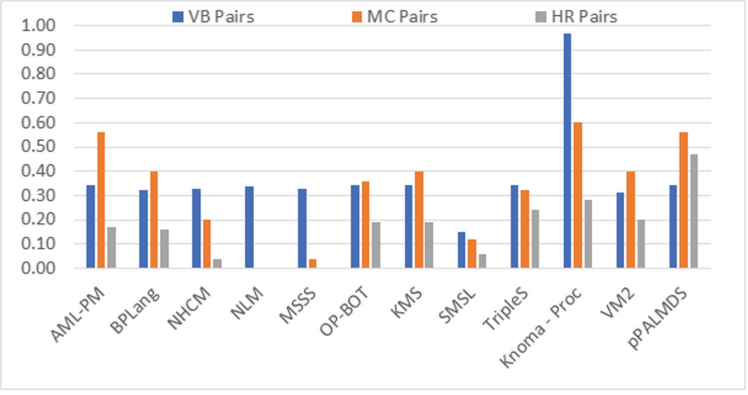

Fig. 7. Performance variation across pairs for AM dataset

5 Conclusion

A plethora of matching techniques have been developed. To evaluate the effectiveness of these techniques researchers typically use PMMC'15 datasets and the performance score of the three performance measures, Precision, Recall and F_1 score. However, our synthesis of the datasets and results of the matching techniques have revealed two issues, (a) the absence of trivial pairs in the datasets artificially inflates the performance scores, (b) the overall performance scores are useful for a surface level evaluation of a technique, however in the presence of diverse corresponding pairs it does not provide the necessary insights about the capabilities of the techniques. For instance, it does not answer important questions, such as how many trivial corresponding pairs are identified and many harder-to-detect corresponding pairs are identified.

In this paper, we address these two issues by enhancing the PMMC'15 datasets. For the enhancements, we have pre-processed the gold standards included in the PMMC'15 datasets and classified the corresponding pairs into three types, depending upon the level of differences in the participating labels. To do that, we have proposed three types of corresponding pairs as well as the criteria for classifying corresponding pairs. The three types are, Verbatim, Modified Copy and Heavy Revision. The enhancements are performed by two independent researchers and conflicts are resolved by a consensus approach. Accordingly, the generated enhanced dataset has 1239 corresponding pairs, including 553 Verbatim, 157 Modified Copy and 529 Heavy Revision pairs.

We further propose that the typically computed overall performance scores should be complemented with the separately computed performance scores of the three types of pairs. The typically computed performance scores are useful for a surface level evaluation of matching techniques and the performance scores of the pairs provides valuable insights about the capabilities of the matching techniques. Hence, the combination of these performance measures are effective tools for a comprehensive evaluation of process model matching techniques.

The enhanced dataset is used for a fair evaluation of three matching techniques. The results reveal that the semantic matching measures do not exhibit any promise for identifying Modified Copy and Heavy Revision Pairs. Hence, highlighting the need for a next generation of matching techniques that can show promising result for the two types of pairs. We have also used the enhanced dataset for the evaluation of the matching that participated in Process Model Contest 2015. Based on a thorough analysis of the results we conclude that pPALMDS is the best matching system. The directions for future work includes, developing the next generation of matching techniques, and a systematic procedure, that can guide the evaluation of matching techniques.

References

1. Kuss, E., Leopold, H., van der Aa, H., Stuckenschmidt, H., Reijers, H.A.: Probabilistic evaluation of process model matching techniques. In: Comyn-Wattiau, I., Tanaka, K., Song, I.-Y., Yamamoto, S., Saeki, M. (eds.) ER 2016. LNCS, vol. 9974, pp. 279–292. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46397-1_22
2. Jabeen, F., Leopold, H., Reijers, Hajo A.: How to make process model matching work better? An analysis of current similarity measures. In: Abramowicz, W. (ed.) BIS 2017. LNBIP, vol. 288, pp. 181–193. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59336-4_13
3. Rodríguez, C., Klinkmüller, C., Weber, I., Daniel, F., Casati, F.: Activity matching with human intelligence. In: La Rosa, M., Loos, P., Pastor, O. (eds.) BPM 2016. LNBIP, vol. 260, pp. 124–140. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45468-9_8
4. Awad, A., Polyvyanyy, A., Weske, M.: Semantic querying of business. Process models. In: Proceedings of the 12th IEEE International Conference on Enterprise Distributed Object Computing Conference (EDOC 2008), pp. 85–94, Munich, Germany (2008)
5. Dumas, M., García-Bañuelos, L., La Rosa, M., Uba, R.: Fast detection of exact clones in business process model repositories. Inf. Syst. **38**(4), 619–633 (2012)
6. La Rosa, M., Dumas, M., Uba, R., Dijkman, R.M.: Business process model merging: an approach to business process consolidation. ACM Trans. Softw. Eng. Methodol. **22**(2), 11–42 (2012)
7. Meilicke, C., Leopold, H., Kuss, E.S., Reijers, H.: Overcoming individual process model matcher weaknesses using ensemble matching. Decis. Support Syst. **100**(1), 15–26 (2017)
8. Antunes, G., et al.: The process model matching contest 2015. In: Kolb, J., Leopold, H., Mendling, J. (eds.) Proceedings of the 6th International Workshop on Enterprise Modelling and Information Systems Architecture (EMISA 2015), Innsbruck, Austria. LNI, pp. 1–29. Springer, Heidelberg (2015)
9. Kuss, E., Leopold, H., Aa, H., Stuckenschmidt Reijers, H.A.: A probabilistic evaluation procedure for process model matching techniques. DKE J. (2018, in press)
10. Sonntag, A., Hake, P., Fettke, P., Loos, P.: An approach for semantic business process model matching using supervised machine learning. In: Proceedings of the 24th European Conference on Information Systems, pp. 1– 12. AIS, Istanbul (2016)
11. Clough, P., Gaizauskas, R., Piao, S., Wilks, Y.: METER: MEasuring TExt Reuse. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002), Philadelphia, USA, pp. 152–159 (2002)
12. Clough, P., Stevenson, M.: Developing a corpus of plagiarized short answers. Lang. Resour. Eval. **45**(1), 5–24 (2011)
13. Sameen, S., Sharjeel, M., Nawab, R.M.A., Rayson, P., Muneer, I.: Measuring short text reuse for the Urdu language. IEEE Access **6**(1), 7412–7421 (2018)
14. Xiao, C., Wang, W., Lin, X., Yu, J.X.: Efficient similarity joins for near duplicate detection. ACM Trans. Database Syst. **36**(3), 1–15 (2011)
15. Miller, A.G.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
16. Patwardhan, S., Banerjee, S., Pedersen, T.: Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics, Maxico City, Mexico, pp. 241–257 (2003)
17. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Comput. Linguist. **32**(1), 13–47 (2006)

18. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning (ICML 1998), Madison, USA, pp. 296–304 (1998)
19. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from a ice cream cone. In: Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC 1986), Toronto, Canada, pp. 24–26 (1986)
20. Niemann, M., Siebenhaar, M., Schulte, S., Steinmetz, R.: Comparison and retrieval of process models using related cluster pairs. Comput. Ind. **63**(2), 168–180 (2012)
21. Sebu, M.L.: Similarity of business process models in a modular design. In: Proceedings of the Applied Computational Intelligence and Informatics (SACI 2016), Timisoara, Romania, pp. 31–36 (2016)