Ranking-based Evaluation of Process Model Matching (Short Paper)

Elena Kuss[1], Henrik Leopold[2], Christian Meilicke[1], and Heiner Stuckenschmidt[1]

[1] Research Group Data and Web Science
University of Mannheim, 68163 Mannheim, Germany
elena|christian|heiner@informatik.uni-mannheim.de
[2] Department of Computer Science
Vrije Universiteit Amsterdam
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
h.leopold@vu.nl

Abstract. Process model matching refers to the automatic detection of semantically equivalent or similar activities between two process models. The output of process model matchers is the basis for many advanced process model analysis techniques and, therefore, must be as accurate as possible. Measuring the performance of process model matchers, however, is a difficult task. On the one hand, it is hard to define which correspondences are actually correct. On the other hand, it is challenging to appropriately take the output of matchers into account, because they often produce confidence values between zero and one. In this paper, we propose the first evaluation procedure for process model matchers that addresses both of these challenges. The core idea is to rank both the computed and the desired correspondences based on their confidence values and compare them using the Spearman's rank correlation coefficient. We perform an in-depth evaluation in which we apply the new evaluation procedure and illustrate how it helps gaining interesting insights.

Keywords: Process Model Matching, Ranking-based Evaluation, Non-binary Gold Standard

1 Introduction

Process models are conceptual models used for a variety of purposes ranging from business process documentation to requirements definition [6]. Process model *matching* is concerned with the automatic identification of semantically equivalent or similar activities between such models. The application scenarios of process model matching are manifold. They include the analysis of model differences [10], harmonization of process model variants [11], and process model search [7]. The challenges associated with the matching task are considerable. Among others, process model matchers must be able to deal with heterogeneous

vocabulary, different levels of granularity, and the fact that typically only a few activities from one model have a corresponding counterpart in the other. In recent years, a significant number of process model matchers have been defined to address these problems (cf. [3, 8, 12, 18, 19]).

One important question that concerns all these matchers is how to evaluate whether they actually perform well. To measure the performance of process model matchers, their final output is compared to a manually annotated gold standard. A key problem in this context is that it is hard to define which correspondences are actually correct. A recently introduced evaluation procedure for process model matchers addresses this problem by introducing the notion of a non-binary gold standard [9]. The idea of a non-binary gold standard is to associate each activity correspondence with a confidence value instead of defining it as correct or incorrect. However, this evaluation procedure still assumes that the output of the matcher is binary. In fact, many matchers produce confidence values that indicate the reliability of the identified correspondences. The transformation of these confidence values into binary values does not only come with the loss of information, but also results in a less accurate assessment of the performance of the matching technique.

In this paper, we therefore introduce the first evaluation procedure for process model matchers that takes the non-binary output of matchers as input and compares it against a non-binary gold standard. To this end, we rank the correspondences produced by the matcher and the gold standard based on their confidence values and compare them using the Spearman's rank correlation coefficient. We perform an in-depth evaluation where we apply the new evaluation procedure and illustrate how it helps in gaining interesting insights.

2 Problem Statement

The goal of evaluation procedures for process model matching is to assess which of the correspondences identified by a matcher are correct. However, there are several problems associated with this task. To illustrate these problems, consider the example depicted in Figure 1. It shows two simplified process models from the Process Model Matching Contest 2015 [2]. Possible correspondences are denoted using gray shades.

Upon closer inspection of the correspondences shown in Figure 1, it becomes clear that some of the correspondences are actually disputable. Consider, for instance, the correspondence between "*Receive online application*" from University 1 and "*Receive application form*" in the process of University 2. On the one hand, we can argue in favor of this correspondence because both activities deal with the receipt of an application document. On the other hand, we can argue that these activities do not correspond to each other because the former relates to an online procedure whereas the second refers to a paper-based application. Similar arguments can be brought forward for the correspondence between "*Invite for interview*" and "*Invite for aptitude test*". We could argue that both activities represent a means to select a promising candidate. However, we can

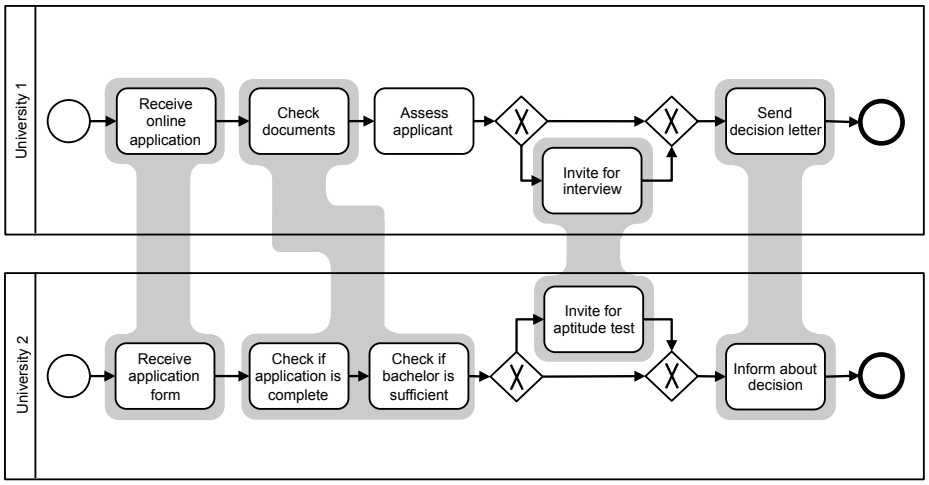

Fig. 1. Two process models and possible correspondences as shown in [9]

also argue that an interview is clearly a different assessment instrument than an aptitude test.

These examples illustrate that it may be hard and, in some cases, even impossible to agree on a single *correct* set of correspondences. Traditional process model matching evaluation procedures, however, assume that such a set of correct correspondences, a so-called *gold standard*, is available. Based on this gold standard, they distinguish between correct and incorrect correspondences generated by a matcher and compute the traditional evaluation metrics precision, recall, and F-measure (cf. [2,4,12,18,19]). Recently, Kuss et al. [9] introduced the notion of a non-binary gold standard. Such a non-binary gold standard assigns a confidence value to correspondences instead of defining them as either correct or incorrect. What this approach, however, still does not take into account is that also matchers often generate confidence values between zero and one. The transformation of these confidence values into binary values for the purpose of evaluating the performance of a matcher with precision, recall, and F-measure does not only come with the loss of information, but also does not result in a fair assessment of the performance.

Recognizing these shortcomings, we use this paper to propose a ranking-based evaluation procedure for process model matching. To this end, we replace the question whether a correspondence is correct with the question in how far the confidence estimated by a matcher resembles the confidence values in the gold standard.

3 Probabilistic Evaluation of Process Model Matching

Given two process models $\mathcal{M}_1$ and $\mathcal{M}_2$, let $\mathcal{G}$ be a non-binary alignment between $\mathcal{M}_1$ and $\mathcal{M}_2$ that represents the manually created gold standard and $\mathcal{A}$ be a

non-binary alignment between $\mathcal{M}_1$ and $\mathcal{M}_2$ that was generated by a matching techniqe. In the following, we show how to compute and use the Spearman's rank correlation coefficient [17] to measure the quality of $\mathcal{A}$ given the manually created gold standard $\mathcal{G}$. Let n be the number of correspondences with a confidence value higher than zero in $\mathcal{G}$ or $\mathcal{A}$, i.e., $n = |\{(a_1, a_2) \in act(\mathcal{M}_1) \times act(\mathcal{M}_1) \mid \mathcal{A}(a_1, a_2) > 0 \vee \mathcal{G}(a_1, a_2) > 0\}|$. To compute the rank correlation, the following steps need to be performed independently for both $\mathcal{G}$ and $\mathcal{A}$.

Normalized Ranks The n correspondences in $\mathcal{G}$ and $\mathcal{A}$ have to be ranked according to their confidence values (in increasing order). This leads to a rank of 1 through n for each correspondence. If there are correspondences with the same confidence value, their ranks are normalized. In these cases, which we refer to as ties, the rank of each correspondence with the same confidence value is given by the arithmetic mean of the ranks occupied by these correspondences.

Correction Term for Ties The number of times each value is observed in the alignment is counted. This is denoted by $t_{\mathcal{A},k}$ with respect to $\mathcal{A}$ and $t_{\mathcal{G},k}$ with respect to $\mathcal{G}$. The index k is used here to refer to the different values (or ranks). As a result of this counting, we obtain $\sum_k t_{\mathcal{A},k} = \sum_k t_{\mathcal{G},k} = n$. In the final formula, we need to use the correction terms $T_{\mathcal{A}} = \sum_k \left((t_{\mathcal{A},k})^3 - t_{\mathcal{A},k} \right)$ and $T_{\mathcal{G}} = \sum_k \left((t_{\mathcal{G},k})^3 - t_{\mathcal{G},k} \right)$.

We can now use the following formula to compute Spearman's rank correlation coefficient, where d_i denotes the difference between the normalized ranks of the i-correspondence from those correspondences that have a positive confidence value in $\mathcal{G}$ or $\mathcal{A}$:

$$\rho = \frac{n^3 - n - \frac{1}{2}T^{\mathcal{G}} - \frac{1}{2}T^{\mathcal{A}} - 6\sum_{i=1}^{n} d_i^2}{\sqrt{(n^3 - n - T^{\mathcal{G}})\,(n^3 - n - T^{\mathcal{A}})}}.$$

4 Evaluation Experiments

In this section, we demonstrate the capabilities of the proposed evaluation procedure by applying it to a set of seven process model matchers. The goal of the evaluation is to show that our evaluation procedure represents a viable alternative to binary evaluation procedures and provides useful analytical insights.

4.1 Setup

To evaluate the proposed evaluation procedure, we applied it to the output of seven matchers that participated in the Process Model Matching Contest 2015 and the Process Model Matching Track at the OAEI 2016. Note that we had to limit our evaluation to seven matchers because not all of the matchers participating in these events provided confidence values for the correspondences they generated.

	nB-nB		B-nB				B-B			
Matcher	R	ρ	R	ProFM	ProP	ProR	R	FM	Prec	Rec
AML	1	**.245**	1	**.424**	.806	.288	1	**.702**	.719	**.685**
Match-SSS	2	.223	5	.314	**.828**	.194	2	.608	**.807**	.487
LogMap	3	.153	2	.418	.680	.302	5	.481	.449	.517
Know-Match-SSS	4	.120	3	.409	.676	.293	3	.544	.513	.578
TripleS	5	-.008	6	.300	.519	.211	4	.485	.487	.483
AML-PM	6	-.266	4	.407	.411	**.404**	6	.385	.269	.672
pPalm-DS	7	-.295	7	.276	.230	.346	7	.253	.162	.578

Table 1. Results for the seven considered matchers from the Process Model Matching Contest 2015 and the OAEI 2016 for three evaluation procedures.

The *data set* that was used in these events consists of 36 model pairs derived from nine process models (referred to as *University Admission* data set), which describe the application procedures for accepting graduate students of nine German universities. The models vary in size and consist between 10 to 44 activities. In the context of both events, the task was to match these models pair-wise. For more details about the data set, we refer the reader to [1, 2].

For the creation of the *non-binary gold standard*, eight individuals were asked to independently create a binary gold standard for the University Admission data set. The resulting eight individually created binary gold standards were merged into a non-binary gold standard. Each correspondence in the non-binary gold standard has an associated non-binary confidence value which represents the share of the eight individual gold standards that contain the respective correspondence.

4.2 Results

As a result of applying our evaluation procedure to the output of the considered seven matchers, we obtained a respective rank correlation coefficient for each matcher. Table 1 gives an overview of the results. It shows the evaluation metrics and the rank (R) for three different evaluation procedures:

- *nB-nB*: The non-binary evaluation procedure introduced in this paper. The performance is captured using the rank correlation coefficient (ρ).
- *B-nB*: The probabilistic evaluation procedure from [9], which compares the binary output of a matcher against a non-binary gold standard. The performance is captured using the probabilistic F-measure (ProFM), probabilistic precision (ProP), and probabilistic recall (ProR).
- *B-B*: The classical evaluation procedure comparing the binary output of a matcher against a binary gold standard. The performance is captured using the F-measure (FM), precision (Prec), and recall (Rec).

The results from Table 1 reveal that there is a rather weak correlation between the output of the matchers and the non-binary gold standard. Three

matchers even have a negative correlation coefficient. This outcome can be explained by the characteristics of the matchers as well as the characteristics of the gold standard. To understand how the characteristics of the matchers can explain this outcome, consider the metrics from the other two evaluation procedures (i.e. B-nB and B-B). All three matchers with a negative correlation coefficient have a particularly low precision, i.e. smaller than 0.5 for the classical and 0.519 for the probabilistic version. Apparently, a negative correlation coefficient primarily relates to a high number of false positives. A notable characteristic of the non-binary gold standard that contributed to the weak correlation is the high number of correspondences with a low support value. The non-binary gold standard contains a total of 831 correspondences, of which about 20% have the lowest rank, i.e. at most one of the eight annotators has voted for them. It is, thus, not surprising that many matchers miss these correspondences. While the penalty for missing them is rather low, the recall values reveal that they also explain the overall correlation coefficient. Abstracting from the absolute values, we see that the correlation coefficient allows us to rank the matchers according to their performance. What is particularly interesting is that the ranking obtained through the evaluation procedure presented in this paper does not always deviate from the ranking we obtain when using the other evaluation procedures. In fact, the matcher AML is always considered to perform best and the matcher pPalm-DS is always considered to be worst.

All in all, the results highlight a major difference of the presented evaluation procedure to existing ones: The confidence of the matcher is taken into account. If a matcher identifies a correspondences that is not part of the gold standard with high certainty, the penalty is much higher than if the certainty is low. This is an important difference to both the B-nB and B-B evaluation procedures where the output of the matcher is considered as zero or one. This particular feature of our evaluation procedure also explains the different ranking in Table 1. Matchers that identify false positives with high certainty receive a bigger penalty than matchers that identify false positives with low certainty.

5 Related Work

To the present day, the evaluation using precision, recall, and F-measure still represents the standard procedure for assesssing the performance of process model matching techniques, see for example the reports of the Process Model Matching Contests [2, 4]. In fact, this also applies to the related fields of schema and ontology matching, which also aim at identifying relations between different conceptual models [14, 16]. However, these fields use a broader range of evaluation metrics to also address the needs related to specific application scenarios (see e.g. [13]).

One of the first evaluation procedures that builds on a non-binary alignment as input has been proposed by Sagi and Gal [15]. They adapt precision and recall metrics in such a way that they can be directly applied to first-line-matching results with non-binary confidence values. Their approach, however, still requires a

binary gold standard as input. In [5], the authors directly compare the confidence values of the matchers to the confidence values of a gold standard. However, the confidence values are not normalized to the same range. As a result, the performance evaluation is quite questionable, since many matchers use largely differing ranges of confidence values. This is a weakness we address with the evaluation procedure proposed in this paper.

6 Conclusion

In this paper, we addressed the problem of how to properly evaluate the quality of process model matchers. Recognizing that binary evaluation procedures based on F-Measure, precision, and recall are insufficient, we introduced the first evaluation procedure for process model matchers that takes the non-binary output of a matcher as input and compares it against a non-binary gold standard. Our evaluation procedure builds on ranking the correspondences produced by the matcher and the gold standard based on their confidence values and comparing them using the Spearman's rank correlation coefficient. The core idea is that the confidence value distribution of the matcher should resemble the confidence value distribution of the gold standard as closely as possible.

To illustrate the usefulness and applicability of our evaluation procedure, we applied it to the output of seven process model matchers that participated in the Process Model Matching Contest 2015 and in the Process Model Matching Track of the OAEI 2016. The results show that our evaluation procedure indeed delivers useful results. While the assessment with respect to the best and the worst performance is congruent with other evaluation procedures, our non-binary procedure also assesses some matchers differently. By considering the confidence values produced by the matchers, it is able to assign a bigger penalty to those matchers that identify false positives with high certainty than to those techniques that identify false positives with little certainty. As a result, the performance of matchers is more accurately assessed.

In future work, we set out to apply the novel evaluation procedure in the context of comparative evaluation experiments. Our goal is to increase the awareness about the necessity to use metrics other than F-Measure, precision, and recall to assess uncertain problems such as process model matching.

References

1. Achichi, M., Cheatham, M., Dragisic, Z., Euzenat, J., Faria, D., Ferrara, A., Flouris, G., Fundulaki, I., Harrow, I., Ivanova, V., Jiménez-Ruiz, E., Kuss, E., Lambrix, P., Leopold, H., Li, H., Meilicke, C., Montanelli, S., Pesquita, C., Saveta, T., Shvaiko, P., Splendiani, A., Stuckenschmidt, H., Todorov, K., Trojahn, C., Zamazal, O.: Results of the ontology alignment evaluation initiative 2016. In: CEUR workshop proceedings. vol. 1766, pp. 73–129. RWTH (2016)
2. Antunes, G. et al.: The process model matching contest 2015. In: 6th International Workshop on Enterprise Modelling and Information Systems Architectures (2015)

3. Cayoglu, U., Oberweis, A., Schoknecht, A., Ullrich, M.: Triple-s: A matching approach for Petri nets on syntactic, semantic and structural level. Tech. rep., Karlsruhe Institute of Technology (KIT) (2013)
4. Cayoglu, U. et al.: The process model matching contest 2013. In: 4th International Workshop on Process Model Collections: Management and Reuse (PMC-MR'13) (2013)
5. Cheatham, M., Hitzler, P.: Conference v2. 0: An uncertain version of the oaei conference benchmark. In: International Semantic Web Conference. pp. 33–48. Springer (2014)
6. Dumas, M., Rosa, M., Mendling, J., Reijers, H.: Fundamentals of Business Process Management. Springer (2013)
7. Jin, T., Wang, J., La Rosa, M., Ter Hofstede, A., Wen, L.: Efficient querying of large process model repositories. Computers in Industry 64(1), 41–49 (2013)
8. Klinkmüller, C., Weber, I., Mendling, J., Leopold, H., Ludwig, A.: Increasing recall of process model matching by improved activity label matching. In: Business Process Management, pp. 211–218. Springer (2013)
9. Kuss, E., Leopold, H., Van der Aa, H., Stuckenschmidt, H., Reijers, H.A.: Probabilistic evaluation of process model matching techniques. In: Conceptual Modeling: 35th International Conference, ER 2016, Gifu, Japan, November 14-17, 2016, Proceedings 35. pp. 279–292. Springer (2016)
10. Küster, J., Gerth, C., Förster, A., Engels, G.: Detecting and resolving process model differences in the absence of a change log. Business Process Management pp. 244–260 (2008)
11. La Rosa, M., Dumas, M., Uba, R., Dijkman, R.: Business process model merging: An approach to business process consolidation. ACM Transactions on Software Engineering and Methodology (TOSEM) 22(2), 11 (2013)
12. Leopold, H., Niepert, M., Weidlich, M., Mendling, J., Dijkman, R., Stuckenschmidt, H.: Probabilistic optimization of semantic process model matching. In: Business Process Management, pp. 319–334. Springer (2012)
13. Mena, E., Kashyap, V., Illarramendi, A., Sheth, A.: Imprecise answers in distributed environments: Estimation of information loss for multi-ontology based query processing. International Journal of Cooperative Information Systems 9(04), 403–425 (2000)
14. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. the VLDB Journal 10(4), 334–350 (2001)
15. Sagi, T., Gal, A.: Non-binary evaluation for schema matching. In: Conceptual Modeling, pp. 477–486. Springer (2012)
16. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. Knowledge and Data Engineering, IEEE Transactions on 25(1), 158–176 (2013)
17. Spearman, C.: The proof and measurement of association between two things. The American journal of psychology 15(1), 72–101 (1904)
18. Weidlich, M., Dijkman, R., Mendling, J.: The ICoP framework: Identification of correspondences between process models. In: Advanced Information Systems Engineering. pp. 483–498. Springer (2010)
19. Weidlich, M., Sheetrit, E., Branco, M.C., Gal, A.: Matching business process models using positional passage-based language models. In: Conceptual Modeling, pp. 130–137. Springer (2013)