Automatic Classification to Matching Patterns for Process Model Matching Evaluation

Elena Kuss and Heiner Stuckenschmidt

University of Mannheim, Artificial Intelligence Group

Abstract. Business process model matching is concerned with the detection of similarities in business process models. To support the progress of process model matching techniques, efficient evaluation strategies are required. State-of-the-art evaluation techniques provide a grading of the evaluated matching techniques. However, they only offer limited information about strength and weaknesses of the individual matching technique. To efficiently evaluate matching systems, it is required to automatically analyze the attributes of the matcher output. In this paper, we propose an evaluation by automatic classification of the alignments to matching patterns. On the one hand, to understand strength and weaknesses of a matching technique. On the other hand, to identify potential for further improvement. Consequently, optimal matching scenarios of a specific matcher can be derived. This further enables tuning of a matcher to specific applications.

Keywords: business process model matching evaluation, matching performance assessment, matching patterns

1 Introduction

Conceptual models, like business process models, are widely used to document a company's operations. Process model matching aims in identifying semantic similarities in business process models. Application scenarios vary strongly for example from process model search in huge data bases [9], the automatic integration within company merges [11] or clone detection [6]. There are many techniques for the automatic identification of similarities in process models [4,13,17,19]. Current process model matching techniques focus on finding similarities between process models by comparing the label-strings of these process models.

Overall, it can be observed that most research efforts are spent in advancing the process matching techniques compared to the advancement of their evaluation.

The process model matching track (PMMT) at the Ontology Alignment Evaluation Initiative (OAEI) [1] and the Process Model Matching Contests (PMMCs) [2,3] apply different metrics to evaluate process model matching techniques. These metrics allow for evaluating the performance of matching techniques through a ranking. Consequently, they only provide limited information about the performance of matching techniques, e.g., detailed information about

Copyright © by the paper's authors. Copying permitted only for private and academic purposes.
In: C. Cabanillas, S. España, S. Farshidi (eds.):
Proceedings of the ER Forum 2017 and the ER 2017 Demo track,
Valencia, Spain, November 6th-9th, 2017,
published at http://ceur-ws.org

the strength and weaknesses of a matching technique. However, the progress of process model matching techniques is strongly influenced by the available evaluation techniques. Important is that the evaluation techniques are "fair" and that they can be computed efficiently, i.e., without intensive manual labor.

In this paper, we propose a conceptually different evaluation approach. The basis of our idea is to group the alignments of the matcher output as well as the reference alignment into different categories. In our example we utilize five different such categories. For each of the category, the widely-used measures precision, recall and F-measure are computed. These metrics for each category enables us then to analyze a matcher's performance in greater detail. This way, we gain insights which matcher performs well on "trivial" correspondences or can also identify "difficult" correspondences. Among the more complex correspondences, we learn for each individual matcher which correspondences can be found and which are challenging.

In the first step of the proposed evaluation method, the correspondences in the reference alignment are assigned to the different categories. The same is done for the matcher output in the second step. Both steps are done automatically – no user input is required. Then, each category is treated as its own matching problem where standard metrics can be applied.

By automatic annotation to matching patterns, the matching problem itself is classified into categories. These categories divide the matching problem into different levels of complexity. On the one hand this helps gaining insights about the complexity of the matching task itself, on the other hand this helps understanding strength and weaknesses of a matcher. Equally important, the categorization helps to identify possibilities for the improvement of a matcher's performance.

Furthermore, the categorization helps understanding the performance of a matcher for specific matching tasks. Sometimes a matcher needs to satisfy different tasks. For example when finding similar process models in a database it may be required to be able to identify a high fraction of correspondences which are semantically identical. However, for some applications it is simply required to compute a high fraction of identical labels. For an efficient evaluation it is required that the evaluation takes the specific application scenarios into account.

The remainder of the paper is organized as follows. In Section 2, the basis of the process model evaluation is defined. Section 3 introduces the matching patterns which are automatically assigned and gives examples to illustrate these patterns. In Section 4, evaluation experiments are obtained and the results of the evaluation by matching patterns are discussed in detail. While Section 5 gives an overview about related work, Section 6 concludes the paper and discusses future work.

2 Problem Description

Definition 1 (Alignment, reference alignment, matcher output, correspondence). *For two process models P_1, P_2 with their activity sets A_1, A_2,*

an alignment is a set consistent of activity pairs (a_1, a_2) *with* $a_1 \in A_1$ *and* $a_2 \in A_2$*. A reference alignment* G *is a subset of all possible alignments, i.e.,* $G \subseteq A_1 \times A_2$*. Similarly, a matcher output* O *is a subset of all possible alignments, i.e.,* $O \subseteq A_1 \times A_2$*. An alignment of the reference alignment is also called a correspondence.*

The task of a matcher is to identify semantically similar alignments, i.e., to identify the correspondences of the reference alignment. Currently, the matching is mostly based on the label strings of the process models to be matched. Because it is rarely possible to reach a perfect matching, i.e., $O = G$, the matcher output needs to be evaluated. There is a whole body of literature on process model matching techniques. They have all in common that the evaluation focuses on grading the evaluated matchers. Thus, the evaluation is designed to provide a matcher's rank within a group of matchers. However, most evaluation methods are not designed to provide a detailed analysis of an individual matcher output. To obtain such detailed information, currently it is required to manually process and interpret the matcher output to identify possibilities for improvement. In contrast, we propose a new evaluation technique which provides such information for an individual matcher output, without the need for manual processing. Our proposed category-dependent evaluation has the following properties:

- informs about the data set, e.g., the complexity of the matching task
- assesses the reference alignment indirectly, e.g., quality and quantity of manual annotations
- identifies characteristics as well as strength and weaknesses of a matcher
- enables to optimize a matcher to specific application scenarios.

By identifying the strength and weaknesses of a matcher, the proposed evaluation technique may aid the progress of matching techniques. This becomes clearer with the experiments in Section 4.

3 Automatic Classification to Matching Patterns

Figure 3 illustrates the conceptual structure of the automatically assigned categories. The matching problem is divided into "trivial" and "non-trivial" alignments. "Trivial" alignments are any alignments which are identical after normalization.

Definition 2 (Normalization). *For the classification, an activity is normalized, if (1) all stop words are removed, (2) stemming has been applied and (3) case sensitivity is ignored.*

Example of stop-words are "of", "for" and "the". Examples for stemming are "checking" and "checks" transformed into "check".

"Non-trivial" alignments are all other alignments. The "non-trivial" category consists of four sub-categories. In the following, we define these categories and give examples.

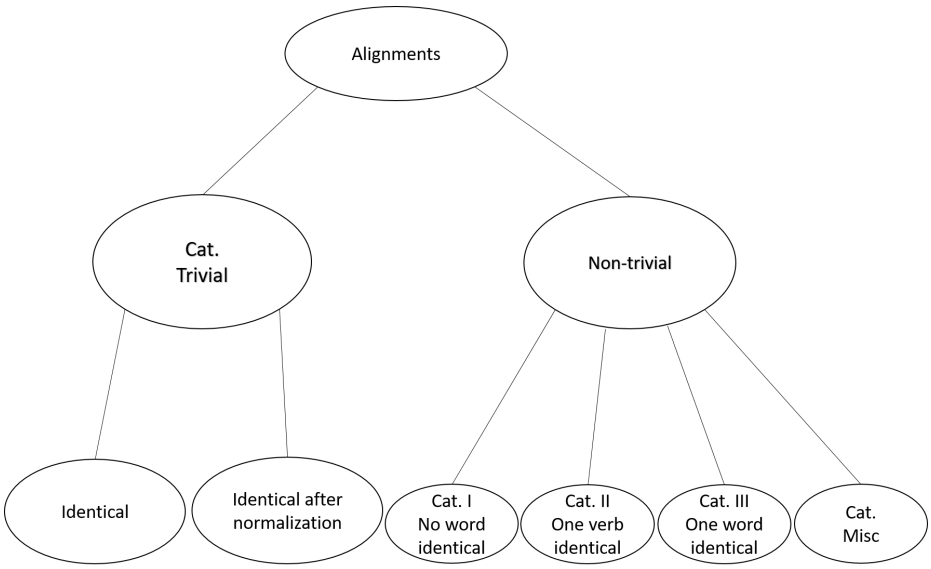

Fig. 1. Structural dependencies of the categories

3.1 The Categories

The core of the evaluation via matching patterns is to automatically [1] assign correspondences of the reference alignment as well as the computed alignments to groups with specific attributes. After the automatically generated classification to one of the categories, the well-known metrics precision, recall and F-measure [14] are calculated for each of theses categories separately. As a consequence, the matching task is divided into groups with specific attributes. In the following, we define the categories and illustrate these with examples from the applied data sets. The categories are chosen to provide a deeper knowledge about specific attributes which are important features of a matching technique. Note that the specific numbering of the categories is not related to the complexity level of the corresponding category.

Definition 3 (Categorization, category). *Let A_1, A_2 be the activity sets for two process models P_1 and P_2. A categorization is a division of all possible alignments into disjoint sets $C(i)$, i.e., $C(i) \subseteq A_1 \times A_2$ for all i with $\cup_i C(i) = A_1 \times A_2$ and $C(i) \cap C(j) = \emptyset$ for all $i \neq j$. Any $C(i)$ is a category.*

In general, the categories should be chosen carefully. It is important that the classification can be assigned automatically. Moreover, too many categories

[1] The implementation of the matching patterns, containing the automatic annotation can be accessed here: https://github.com/kristiankolthoff/PMMC-Evaluator/tree/master/src/main/java/de/unima/ki/pmmc/evaluator/annotator

lead to too few alignments in each category; too few categories lack information. We propose the following categories (the examples are taken from the reference alignment of the data sets of the PMMC 2015):

Category "trivial": This category contains alignments which are identical after normalization.

Category I "no word identical": Alignments which have no word in common after normalization are assigned to this category. Examples:
- **Ex. 1:** *Evaluate – Assessment of application*
- **Ex. 2:** *Hand application over to examining board – Send documents to selection committee*
  [The stop word "to" is ignored and not counted as an identical word.]
- **Ex. 3:** *Talk to applicant – Do the interview*
- **Ex. 4:** *Shipping – Delivery and Transportation Preparation*
- **Ex. 5:** *Shipment – Transportation Planning and Processing*

Category II "one verb identical": Alignments which are assigned to this category have exactly one identical verb after normalization. No other words are identical. Examples:
- **Ex. 6:** ***Send*** *documents by post –* ***Send*** *all the requirements to the secretarial office for students*
- **Ex. 7:** ***Wait*** *for results –* ***Waiting*** *for response*
  [This example illustrates two specific characteristics: the verb is normalized (stemming), the stop word (in this case "for") is ignored.]
- **Ex. 8:** ***Send*** *acceptance –* ***Send*** *commitment*
- **Ex. 9:** ***Check*** *data –* ***Check*** *documents*

Category III "one word identical": This category consists of alignments which have exactly one word (but not a verb) in common after normalization. Examples:
- **Ex. 10:** *Talk to* ***applicant*** *– Appoint* ***applicant***
- **Ex. 11:** *Hand* ***application*** *over to examining board – Send* ***application*** *to selection committee*
  [In this example the stop word "to" is ignored.]
- **Ex. 12:** *Apply* ***online*** *– Fill in* ***online*** *form of application*
- **Ex. 13:** ***Invoice*** *approval –* ***Invoice*** *Verification*

Category "misc": All other alignments, which cannot be assigned to one of the above categories. Examples:
- **Ex. 14:** ***Send application*** *–* ***Send application*** *form and documents*
- **Ex. 15:** ***Send*** *documents to* ***selection committee*** *–* ***Send*** *application to* ***selection committee***
- **Ex. 16:** ***Receiving*** *the written* ***applications*** *–* ***Receive application***
- **Ex. 17:** ***Time Sheet*** *Approval –* ***Time Sheet*** *Permit*

The described categories vary strongly in their level of complexity. In the following, we discuss each category with increasing complexity level of the categories. The Cat. trivial contains identical labels after normalization. Only basic syntactical matching techniques are required to identify such correspondences.

The Cat. misc contains all alignments which cannot be assigned to one of the other categories. Thus, it contains only alignments which share more than one identical word. Therefore this category is a category with less complex alignments compared to Cat. I through Cat. III.

Cat. II and Cat. III have a rather high complexity level, since these categories have just one word / one verb in common. Both categories can further indicate if a matcher produces already a high fraction of alignments if one word or the verb between two labels are identical. Cat. I, however, is the most complex category among the introduced categories, since these alignments have no word in common. They have no syntactical overlap. Consequently these alignments just have a semantic connection, like this is the case for synonyms. To identify alignments from this category correctly, a matcher requires advanced semantic knowledge. Note that each alignment is assigned to exactly one of these categories exclusively, i.e., the alignments cannot be assigned to several categories.

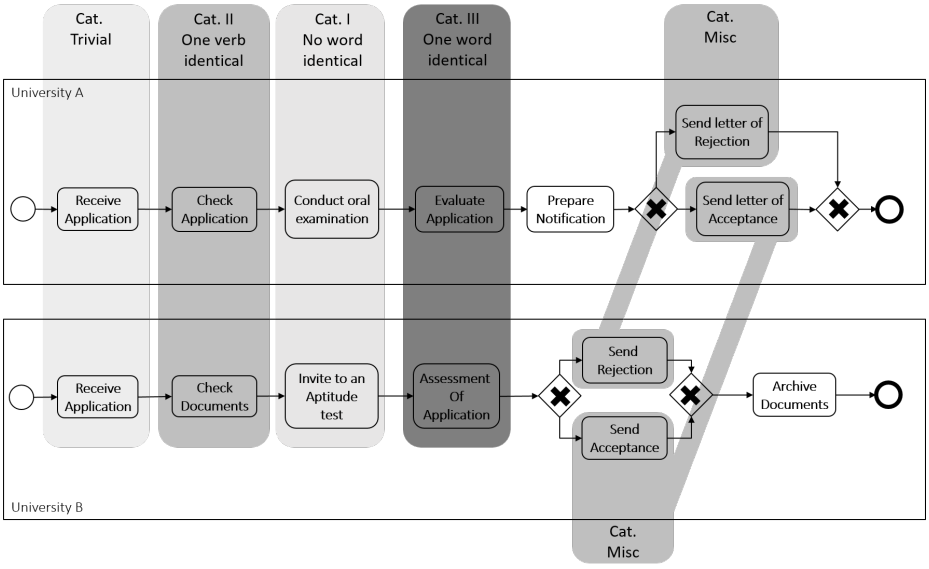

Fig. 2. Example of a categorized reference alignment

Figure 2 illustrates a simplified example of a reference alignment which is assigned to the above described categories. The figure shows two example process models, which illustrate the application process of Master students at two universities. A matcher's task is to identify correspondences of one process model in

the other process model. The correspondences of this matching task are marked with different gray scales for each of the introduced categories above. For each introduced category there is one example in the figure. However, the Cat. misc is illustrated with two examples of the reference alignment. Note that the illustrated figure is an example of a reference alignment. For the evaluation procedure also the alignments computed by a matcher are classified to the matching patterns.

3.2 Metrics for the Categories

Definition 4 (category-dependent Precision, category-dependent Recall, category-dependent F-measure). *Let set G be the reference alignment, O be the matcher output and $C(i)$ be the categories. The set $G(i)$ is the collection of reference alignments assigned to category $C(i)$, i.e., $G(i) = G \cap C(i)$ for all i. Similarly, $O(i)$ is the collection of correspondences computed by a matcher and assigned to category $C(i)$; i.e., $O(i) = O \cap C(i)$ for all i.*

The category-dependent Precision, $cP(i)$, is defined as

$$cP(i) = \frac{|G(i) \cap O(i)|}{|O(i)|}$$

and the category-dependent Recall, $cR(i)$, is given by

$$cR(i) = \frac{|G(i) \cap O(i)|}{|G(i)|} \ .$$

The category-dependent F-measure, $cFM(i)$, is then

$$cFM(i) = 2 \cdot \frac{cP(i) \cdot cR(i)}{cP(i) + cR(i)} \ .$$

Precision is the fraction of correctly computed alignments to all computed alignments. Recall is the fraction of correctly computed alignments to all correct correspondences (with respect to the reference alignment). Both precision and recall are values between 0.0 and 1.0. A precision of 1.0 means that all computed correspondences are contained in the reference alignment, i.e. $O(i) \subseteq G(i)$. In contrast, a recall of 1.0 means that all correspondences of the reference alignment are computed, i.e. $G(i) \subseteq O(i)$. The F-measure is the harmonic mean of precision and recall. All alignments in the reference alignment as well as the matcher output are assigned to exactly one category exclusively, i.e. there is no overlap between these categories. After this, precision, recall and F-measure of the alignments of each category are calculated. That means, the categories are evaluated separately and independently.

4 Experiments

To illustrate the insights which can be obtained by the proposed matching evaluation, we apply the introduced technique to the data sets and participating

matchers of the Business Process Model Matching Contest 2015 and the Process Model Matching Track at the OAEI 2016 with its participating matching techniques.

4.1 Setting

In the following, the data sets are described. More details about the participating matchers and the data sets can be found at [1,2].

Birth Registration Data set (PMMC 2015) The birth registration data set contains 36 business process model pairs, which model the process of registering a new born child in Russia, South Africa, Germany and the Netherlands. The models are modeled as Petri-nets.

Asset Management Data set (PMMC 2015) This data set consists of 36 model pairs of a SAP Reference Model collection which describe processes in the area of finance and accounting. These models are event-driven process chains (EPCs).

University Admission Data set (PMMC 2015 / PMMT 2016) This data set contains 36 process model pairs, derived from 9 models from German universities, which model the application process of Master students.

4.2 Results

Below, we present the results of the automatic generated matching patterns. The matching patterns are assigned automatically to the reference alignment, as well as to the matcher output of the matchers which participated in the PMMC 2015 and the PMMT at the OAEI 2016. Then category-dependent precision, recall and F-measure are computed for each category separately. After application of the matching patterns to the reference alignment as well as to the alignments computed by the matchers, the following results are computed.

Computational results Tables 1-3 illustrate the results for each data set. The first column provides a list of all participating matchers. They are listed in alphabetic order. In the second column, the F-measure (FM) over all matching patterns is given as the micro value, i.e. it is computed over all test cases. The remaining columns provide the category-dependent precision (cP), recall (cR) and F-measure (cFM) for each matcher in each category. cP, cR and cFM are macro values, independently computed for each of the matching patterns. For each category, the tables further show in the heading the fraction of correspondences from the whole data set as well as the total number of correspondences of a category in the reference alignment. The best three matchers are highlighted in each category. One central observation is the distribution of the correspondences

Approach	FM	Cat. trivial [44.3%][103]			Cat. I no word iden. [29.3%][68]			Cat. II one verb iden. [11.6%][27]			Cat. III one word iden. [7.3%][17]			Cat. misc [7.3%][17]		
		cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM
AML	**.698**	.844	.959	.862	**.952**	**.595**	**.623**	**.833**	.300	**.311**	**.667**	.372	**.344**	**.157**	.576	**.183**
AML-PM	.385	.844	.963	.864	.458	.397	.334	.187	**.633**	.217	.045	**.500**	.069	.112	**.970**	.151
BPLangM	.397	**.939**	.816	.864	–	–	–	.462	.344	.262	.152	**.526**	.175	.084	.348	.094
DKP	.538	.844	.968	.867	.267	.048	.070	–	–	–	–	–	–	.136	.227	.099
DKP-lite	.534	.844	.968	.867	–	–	–	–	–	–	–	–	–	.136	.227	.099
KnoMa-Proc	.394	.833	.931	.845	–	–	–	.078	.133	.067	.068	.346	.092	.052	.409	.066
KMSSS	.544	.846	**1.0**	**.883**	.450	.172	.151	**.500**	.289	.251	.357	.205	.164	**.142**	.636	.152
LogMap	.481	.844	.978	.872	–	–	–	.467	.167	.127	.082	.372	.094	.092	.530	.110
MSSS	.608	.844	.968	.867	.500	.069	.057	**.833**	**.500**	**.489**	–	–	–	.143	.091	.083
OPBOT	.601	**.978**	.706	.774	.713	**.468**	**.433**	.562	.322	.290	.432	**.500**	**.333**	.128	.530	**.164**
pPalm-DS	.253	.843	.986	.874	–	–	–	.053	.344	.072	.029	.410	.046	.062	**.939**	.086
RMM-NHCM	**.668**	**.954**	.930	**.928**	**.821**	.374	.397	.452	**.456**	**.292**	**.550**	.372	**.302**	.178	.439	**.166**
RMM-NLM	**.636**	.843	**1.0**	**.881**	.486	.324	.303	–	–	–	–	–	–	**1.0**	.091	.091
RMM-SMSL	.543	.844	.912	.839	**.778**	**.423**	**.439**	.152	.311	.121	–	–	–	.087	.121	.058
RMM-VM2	.293	.825	.767	.759	–	–	–	.044	.367	.065	.040	.372	.058	.081	**.742**	.110
TripleS	.485	.843	**1.0**	**.881**	–	–	–	.077	.156	.072	**.625**	.179	.185	.025	.121	.029

Table 1. Results of University Admission data set (PMMC 2015 and PMMT 2016)

Approach	FM	Cat. trivial [45.9%][102]			Cat. I no word iden. [34.2%][76]			Cat. II one verb iden. [0.9%][2]			Cat. III one word iden. [8.1%][18]			Cat. misc [10.8%][24]		
		cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM
AML-PM	**.677**	**.996**	**1.0**	**.998**	**1.0**	**.059**	**.059**	**.667**	**1.0**	**.667**	.231	**.396**	**.219**	.571	**.742**	**.565**
BPLangM	.646	**.996**	.951	.970	**.300**	**.176**	**.149**	**1.0**	.500	.500	.200	.354	**.161**	**.646**	.706	**.528**
KnoMa-Proc	.355	.251	.968	.367	–	–	–	.083	**1.0**	.119	.100	.062	.042	.220	.570	.237
KMSSS	.579	**.996**	**1.0**	**.998**	–	–	–	–	–	–	**.333**	.062	.067	.342	.552	.297
MSSS	.619	**.996**	**1.0**	**.998**	–	–	–	–	–	–	–	–	–	.417	.127	.119
OPBOT	.639	**.996**	**1.0**	**.998**	.250	.026	.033	.500	**1.0**	.500	.286	**.469**	**.252**	.640	**.891**	**.653**
pPalm-DS	.474	**.996**	**1.0**	**.998**	–	–	–	**1.0**	**1.0**	**1.0**	.243	.312	**.161**	.301	**.909**	.333
RMM-NHCM	**.661**	**.996**	**1.0**	**.998**	–	–	–	–	–	–	**.500**	.062	.083	**.667**	.303	.290
RMM-NLM	**.653**	**.996**	**1.0**	**.998**	–	–	–	–	–	–	–	–	–	**1.0**	.194	.217
RMM-SMSL	.354	.990	.582	.659	–	–	–	–	–	–	**.333**	.177	.144	.333	.109	.089
RMM-VM2	.603	**.996**	.962	.976	**1.0**	**.059**	**.059**	**.667**	**1.0**	**.667**	.131	**.417**	.126	.450	.612	.418
TripleS	.578	**.996**	**1.0**	**.998**	–	–	–	–	–	–	.111	.062	.048	.372	.633	.324

Table 2. Results of Asset Management data set (PMMC 2015)

in the reference alignments. This aids in understanding the complexity level of the applied data sets.

Table 1 illustrates the results for the University Admission data set (including the participants of the PMMT at the OAEI 2016). As can be observed, the University Admission data set consists of a very high fraction of trivial correspondences. Almost half of the correspondences (44,3%) in the reference alignment are trivial correspondences. It can further be observed, that most matchers focus on identifying trivial correspondences. Just few matchers can identify a reasonable number of complex correspondences. Similar behavior can be observed for the Asset Management data set in Table 2 with 45,9% trivial correspondences. Again, most matchers focus on identifying these trivial correspondences. No matcher can achieve good results for Cat. I. For Cat. II and Cat. III there is a similar picture. However, for the Asset Management data set, the number of correspondences in Cat. II is too low to draw meaningful conclusions.

For the Birth Registration data set (Table 3), there is a different observation. 75% of all correspondences are correspondences of Cat. I. This shows, that this

Approach	FM	Cat. trivial [4.5%][26]			Cat. I no word iden. [75.0%][437]			Cat. II one verb iden. [1.5%][9]			Cat. III one word iden. [9.9%][58]			Cat. misc [9.1%][53]		
		cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM
AML-PM	.392	.190	.792	.239	.386	.329	.329	.071	.048	.036	.496	.336	.308	.772	.366	.382
BPLangM	.418	.891	.594	.562	.517	.254	.314	**.500**	**.333**	**.250**	**.554**	.346	**.340**	.742	.253	.295
KnoMa-Proc	.262	.563	.698	.509	.215	.279	.229	.200	.190	.100	.130	.130	.082	.519	.342	.300
KMSSS	.385	.908	.688	.701	**.791**	.239	.308	–	–	–	.450	.148	.143	**.773**	.352	.355
MSSS	.332	**1.0**	.667	.696	**.973**	.174	.243	–	–	–	**.583**	.130	.119	**1.0**	.144	.158
OPBOT	**.565**	.882	**.854**	**.831**	.676	**.422**	**.483**	**.250**	**.286**	**.152**	**.714**	**.485**	**.470**	.688	**.451**	**.444**
pPalm-DS	**.459**	.894	**.875**	**.871**	.454	**.356**	**.354**	.100	**.286**	.111	.452	**.426**	**.335**	.706	**.587**	**.504**
RMM-NHCM	**.456**	**.923**	.698	.717	.717	.319	**.389**	**.400**	.190	**.167**	.552	.262	.261	.623	.292	.283
RMM-NLM	.309	**1.0**	.443	.487	**.952**	.163	.223	–	–	–	.389	.130	.119	**1.0**	.154	.174
RMM-SMSL	.384	.667	.292	.307	.506	.329	.346	.095	**.286**	.111	.304	.194	.147	.633	.390	.353
RMM-VM2	.433	.894	**.854**	**.805**	.391	**.339**	.339	.050	.190	.056	.413	**.432**	**.335**	.652	**.470**	**.469**
TripleS	.384	.445	.719	.450	.651	.268	.309	–	–	–	.433	.142	.131	.679	.367	.361

Table 3. Results of Birth Registration data set (PMMC 2015)

data set is by far the most complex of these three data sets. Similar to the Asset Management data set, only a low fraction of correspondences of the reference alignment have only the verb in common (Cat. II). Furthermore, it can be observed that many matchers fail in identifying the trivial correspondences of this data set. One explanation may be the reference alignment for this data set because we find that the reference alignment of the Birth Registration data set does not fully cover all trivial correspondences or contains wrong trivial alignments. Another observation is that matcher need to take structural dependencies into account, to differentiate between wrong and correct trivial alignments.

The Cat. trivial of the Asset Management and the Birth Registration data sets only contains correspondences which are exactly identical without any normalization. This is not the case for the University Admission data set. Therefore, we further distinguish between the kind of trivial correspondences, i.e. if these correspondences are "identical" or "identical after normalization"; see Figure 3. We find that in the University Admission data set, about 7% of Cat. trivial consists of correspondences which are trivial after normalization, like stemming. This is a very small fraction and illustrates that for the detection of most correspondences of this category not even a normalization is required. However, the sub-division of Cat. trivial, helps to understand if matchers are able to detect trivial correspondences, which require normalization.

Observations and Findings With the evaluation through matching patterns it is possible to identify characteristics, strength and weaknesses of a matcher. The results clearly show that most matchers focus on finding correspondences with low complexity, i.e., Cat. trivial and Cat. misc. The matchers clearly lack identifying complex correspondences. This is especially evident for the Asset Management data set which contains special technical terms. For detecting non-trivial correspondences, a matching technique requires knowledge about these terms. It can be observed that the matcher BPLangMatch, in contrast to the other matchers, is able to identify difficult correspondences of this specific data set. The matcher AML achieves very good results for Cat. I (cFM of 0.623) at

the University Admission data set. With this high performance for this category, it can be expected that the matcher AML would achieve a high performance on the Birth Registration data set. However, since this matcher did not participate in the PMMC 2015, the results for this data set are not available. Moreover, the matcher OPBOT achieves considerably good results for Cat. I in the Admission data set. Therefore it is not surprising that this matcher reaches the best F-measure on the Birth Registration data set.

Approach	University Admission										Asset Management										Birth Registration									
	trivial		I		II		III		misc		trivial		I		II		III		misc		trivial		I		II		III		misc	
	[103]		[68]		[27]		[17]		[17]		[102]		[76]		[2]		[18]		[24]		[26]		[437]		[9]		[58]		[53]	
	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
AML	12	6	2	27	1	22	3	11	45	8																				
AML-PM	12	5	32	48	60	12	113	10	117	1	1	0	0	75	2	0	19	11	13	4	48	5	203	287	9	8	24	38	6	32
BPLangM	6	24	37	68	8	20	55	9	70	10	1	5	15	71	0	1	16	13	8	7	2	8	72	311	3	6	16	37	6	41
DKP	12	4	19	60	0	27	0	17	36	14																				
DKP-lite	12	4	0	68	0	27	0	17	36	14																				
KnoMa-Proc	12	10	1	68	24	23	41	12	134	9	209	7	0	76	13	0	15	17	69	8	12	7	463	306	10	7	35	53	15	37
KM-SSS	12	0	16	61	7	18	9	13	83	6	1	0	0	76	0	2	3	17	61	10	2	7	23	326	1	9	7	52	4	41
LogMap	12	2	0	68	4	23	39	11	92	8																				
Match-SSS	12	4	4	64	2	17	0	17	9	17	1	0	3	76	0	2	0	18	8	21	0	8	6	347	1	9	3	53	0	48
OPBOT	3	21	11	32	8	21	12	10	60	8	1	0	18	74	3	0	27	9	21	2	2	4	74	248	5	5	17	29	12	24
pPalm-DS	13	3	5	68	143	15	317	10	216	2	1	0	4	76	0	0	35	13	163	1	4	3	181	274	10	7	32	34	17	19
RMM-NHCM	8	3	7	42	13	16	5	11	36	9	1	0	0	76	0	2	1	17	3	15	1	7	49	295	3	7	12	43	8	37
RMM-NLM	13	0	25	45	0	27	0	17	0	17	1	0	0	76	0	2	0	18	0	18	0	13	10	352	1	9	7	53	0	46
RMM-SMSL	11	17	6	34	59	15	8	17	43	15	1	56	0	76	0	2	14	14	5	22	4	16	130	291	13	7	19	50	8	39
RMM-VM2	8	20	2	68	114	20	169	11	104	5	1	7	0	75	1	0	39	10	16	9	3	4	190	280	21	7	32	31	13	28
TripleS	13	0	0	68	52	22	2	14	51	16	1	0	0	76	0	2	19	16	56	7	25	6	60	313	1	9	10	52	7	40

Table 4. False-positive (FP) and false-negative (FN) alignments for the three data sets and all matchers, assigned to the categories.

However, from the three different data sets, it seems that the matchers are optimized to the specific data sets. For example, while the matchers focus on finding correspondences from Cat. trivial in the University Admission data set and Asset Management data set, in contrast, at the Birth Registration data set matchers aim at identifying correspondences from Cat. I. This can also be observed by the number of false-positive and false-negative alignments for each category (Table 4). The matchers compute a high number of false-positive alignments in Cat. I for the Birth Registration data set, i.e. the matchers aim at identifying correspondences from this category. For the Asset Management data set, however, most matchers do not compute alignments from Cat. I at all. This can be explained by the fact that Cat. I is on the one hand, the most difficult category, but on the other hand, to succeed at the Birth Registration data set it is necessary to compute correspondences from this category. The reason is the very high fraction of correspondences on the whole Birth Registration data set for Cat. I (about 75%). The classification of the false-positives and false-negatives into the categories allows a more fine-grained understanding about a matcher's performance. It enables to directly identify where sources for errors of the matchers are. For example the matcher KnoMa-Proc computes a very high number of false-positive alignments in Cat. trivial for the Asset Management data set. The matcher RMM-SMSL misses many trivial correspondences (56) from the Asset management data set.

5 Related work

State-of-the-art evaluation techniques mostly rely on precision, recall and F-measure for evaluation of process model matching techniques. In [10] these metrics are extended to probabilistic versions with a non-binary reference alignment. This non-binary reference alignment contains support values obtained by the number of annotators which identify a correspondence. This extension allows a deeper understanding about a matcher's performance, since it helps to understand if matchers focus on identifying correspondences with high support values (mostly obvious correspondences) or if they also can identify arguable correspondences. Furthermore, [18] propose a prediction of the performance of process matching techniques.

In ontology matching, more research has been dedicated to evaluation strategies [5,7,15,16,20]. Euzenat, for example, introduced semantic precision and recall. This extension for ontology matching techniques allows to differentiate if the computed correspondences are related, by taking the ontological structure into account. As a consequence, deductible alignments are evaluated. Although it better reflects the quality of the computed alignments, it does not provide information about the structure of correspondences a matcher can identify and thus does not provide information about specific strength and weaknesses of matching techniques.

In schema matching, [12] propose synthetic scenarios to tune a schema matcher to specific applications. The annual Ontology Alignment Evaluation Initiatives apply synthetic data sets which allow to test matching systems on specific characteristics [1]. However, these synthetic data sets are artificially generated test cases which rely on the same resources as most matchers rely on, e.g. WordNet. Therefore experts generated synonyms manually, but the manual synonym generation comes with high efforts.

In [8], the authors explain that it is not suitable to apply matchers on synthetic scenarios, since these scenarios are too artificial. It is not clear if matchers have a similar performance on real-world data. The matching patterns, proposed in this paper, offer the same features as synthetic data sets, but with real-world data.

6 Conclusion

In this paper we propose a conceptually new approach to divide the matching task as well as the matcher output into patterns with specific attributes. The proposed evaluation via matching patterns provides an in-depth evaluation about a matcher's performance, including specific strength and weaknesses. Therefore it allows for an application dependent evaluation, as the evaluation procedure can aid in improving matching techniques to obtain desired attributes. The evaluation procedure further is an efficient way to automatically process the matcher output. It can help to understand what kind of false-positive and false-negative alignments matchers generate and therefore enables for an quantitative as well

as qualitative analysis. The approach can be extended by different matching patterns, the only limitation is that they can be assigned automatically. Furthermore, all standard metrics can be applied. In future work we aim to use the matching patterns for a prediction of the matcher's performance for specific data sets. We further plan to apply the evaluation technique to the participating matchers of the next Process Model Matching Contest.

Acknowledgments. We thank Kristian Kolthoff for his work on implementing the automatic assignment of the problem categories.

References

1. Manel Achichi, Michelle Cheatham, Zlatan Dragisic, Jérôme Euzenat, Daniel Faria, Alfio Ferrara, Giorgos Flouris, Irini Fundulaki, Ian Harrow, Valentina Ivanova, et al. Results of the ontology alignment evaluation initiative 2016. In *11th ISWC workshop on ontology matching (OM)*, pages 73–129, 2016.
2. Goncalo Antunes, Marzieh Bakhshandeh, Jose Borbinha, Joao Cardoso, Sharam Dadashnia, Chiara Di Francescomarino, Mauro Dragoni, Peter Fettke, Avigdor Gal, Chiara Ghidini, et al. The process model matching contest 2015. *GI-Edition/Proceedings: Lecture notes in informatics*, 248:127–155, 2015.
3. Ugur Cayoglu, Remco Dijkman, Marlon Dumas, Peter Fettke, Luciano García-Bañuelos, Philip Hake, Christopher Klinkmüller, Henrik Leopold, André Ludwig, Peter Loos, et al. Report: The process model matching contest 2013. In *International Conference on Business Process Management*, pages 442–463. Springer, 2013.
4. Remco Dijkman, Marlon Dumas, and Luciano García-Bañuelos. Graph matching algorithms for business process model similarity search. In *International Conference on Business Process Management*, pages 48–63. Springer, 2009.
5. Marc Ehrig and Jérôme Euzenat. Relaxed precision and recall for ontology matching. In *Proc. K-Cap 2005 workshop on Integrating ontology*, pages 25–32. No commercial editor., 2005.
6. Chathura C Ekanayake, Marlon Dumas, Luciano García-Bañuelos, Marcello La Rosa, and Arthur HM ter Hofstede. Approximate clone detection in repositories of business process models. In *International Conference on Business Process Management*, pages 302–318. Springer, 2012.
7. Jérôme Euzenat. Semantic precision and recall for ontology alignment evaluation. In *IJCAI*, pages 348–353, 2007.
8. Jérôme Euzenat, Christian Meilicke, Heiner Stuckenschmidt, Pavel Shvaiko, and Cássia Trojahn. Ontology alignment evaluation initiative: six years of experience. In *Journal on data semantics XV*, pages 158–192. Springer, 2011.
9. Tao Jin, Jianmin Wang, Marcello La Rosa, Arthur Ter Hofstede, and Lijie Wen. Efficient querying of large process model repositories. *Computers in Industry*, 64(1):41–49, 2013.
10. Elena Kuss, Henrik Leopold, Han Van der Aa, Heiner Stuckenschmidt, and Hajo A Reijers. Probabilistic evaluation of process model matching techniques. In *Conceptual Modeling: 35th International Conference, ER 2016, Gifu, Japan, November 14-17, 2016, Proceedings 35*, pages 279–292. Springer, 2016.

11. Marcello La Rosa, Marlon Dumas, Reina Uba, and Remco Dijkman. Business process model merging: An approach to business process consolidation. *ACM Trans. Softw. Eng. Methodol.*, 22(2):11:1–11:42, March 2013.
12. Yoonkyong Lee, Mayssam Sayyadian, AnHai Doan, and Arnon S Rosenthal. etuner: tuning schema matching software using synthetic scenarios. *The VLDB JournalThe International Journal on Very Large Data Bases*, 16(1):97–122, 2007.
13. Henrik Leopold, Mathias Niepert, Matthias Weidlich, Jan Mendling, Remco Dijkman, and Heiner Stuckenschmidt. Probabilistic optimization of semantic process model matching. In *International Conference on Business Process Management*, pages 319–334. Springer, 2012.
14. Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.
15. Hela Sfar, Anja Habacha Chaibi, Amel Bouzeghoub, and Henda Ben Ghezala. Gold standard based evaluation of ontology learning techniques. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 339–346. ACM, 2016.
16. Pavel Shvaiko and Jérôme Euzenat. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176, 2013.
17. Sergey Smirnov, Remco Dijkman, Jan Mendling, and Mathias Weske. Meronymy-based aggregation of activities in business process models. *Conceptual Modeling–ER 2010*, pages 1–14, 2010.
18. Matthias Weidlich, Tomer Sagi, Henrik Leopold, Avigdor Gal, and Jan Mendling. Predicting the quality of process model matching. In *Business Process Management*, pages 203–210. Springer, 2013.
19. Matthias Weidlich, Eitam Sheetrit, Moisés C Branco, and Avigdor Gal. Matching business process models using positional passage-based language models. In *International Conference on Conceptual Modeling*, pages 130–137. Springer, 2013.
20. Elias Zavitsanos, George Paliouras, and George A Vouros. Gold standard evaluation of ontology learning methods through ontology transformation and alignment. *IEEE Transactions on Knowledge and Data Engineering*, 23(11):1635–1648, 2011.