Multilingual Ontology Matching Evaluation – A First Report on Using MultiFarm

Christian Meilicke[1(✉)], Cássia Trojahn[2], Ondřej Šváb-Zamazal[3], and Dominique Ritze[1]

[1] University of Mannheim, Mannheim, Germany
christian@informatik.uni-mannheim.de
[2] INRIA and LIG, Grenoble, France
[3] University of Economics, Prague, Czech Republic

Abstract. This paper reports on the first usage of the MultiFarm dataset for evaluating ontology matching systems. This dataset has been designed as a comprehensive benchmark for multilingual ontology matching. In a first set of experiments, we analyze how state-of-the-art matching systems – not particularly designed for the task of multilingual ontology matching – perform on this dataset. These experiments show the hardness of MultiFarm and result in baselines for any algorithm specifically designed for multilingual ontology matching. We continue with a second set of experiments, where we analyze three systems that have been extended with specific strategies to solve the multilingual matching problem. This paper allows us to draw relevant conclusions for both multilingual ontology matching and ontology matching evaluation in general.

1 Introduction

Ontology matching is the task of finding correspondences that link concepts, properties or instances between two ontologies. Different approaches have been proposed for performing this task. They can be classified along the ontology features that are taken into account (labels, structures, instances, semantics) or with regard to the kind of disciplines they belong to (e.g., statistics, combinatorial, semantics, linguistics, machine learning, or data analysis) [4,8,12].

With the aim of establishing a systematic evaluation of matching systems, the Ontology Alignment Evaluation Initiative (OAEI)[1] [3] has been carried out over the last years. It is an annual evaluation campaign that offers datasets, from different domains, organized by different groups of researchers. However, most of the OAEI datasets have been focused on monolingual tasks. A detailed definition on multilingual and cross-lingual ontology matching tasks can be found in [14]. The multilingual datasets so far available contain single pairs of languages, as the MLDirectory dataset,[2] which consists of website directories in English and Japanese; and the VLCR dataset,[3] that aims at matching the thesaurus of the

[1] http://oaei.ontologymatching.org/.
[2] http://oaei.ontologymatching.org/2008/mldirectory/.
[3] http://www.cs.vu.nl/~laurah/oaei/2009/.

E. Simperl et al. (Eds.): ESWC 2012 Satellite Events, LNCS 7540, pp. 132–147, 2015.
DOI: 10.1007/978-3-662-46641-4_10

Netherlands Institute for Sound and Vision, written in Dutch, to the WordNet and DBpedia, in English. Furthermore, these datasets contain only partial reference alignments or are not fully open. Thus, they are not suitable for an extensive evaluation.

For overcoming the lack of a comprehensive benchmark for multilingual ontology, the MultiFarm dataset has been designed. This dataset is based on the OntoFarm [16] dataset, which has been used successfully in OAEI in the Conference track. MultiFarm is composed of a set of seven ontologies translated in eight different languages and the complete corresponding alignments between these ontologies.

In this paper, we report on the first usage of MultiFarm for multilingual ontology matching evaluation. In [10], we have deeply discussed the design of MultiFarm, focusing on its multilingual features and the specificities of the translation process, with a very preliminary report on its evaluation. Here, we extend this preliminary evaluation and provide a deep discussion on the performance of matching systems. Our evaluation is based on a representative subset of MultiFarm and a set of state-of-the-art matching systems participating in OAEI campaigns. Most of these systems have not particularly been designed for matching ontologies described in different languages. This hold for those systems participating in OAEI 2011. For these systems we have omitted testcases in which Russian and Chinese languages were involved. We also included three participants of OAEI 2011.5 that use specific multilingual components. These systems use basic translation components that are executed prior to the matching process itself. Here we also included Russian and Chinese testcases. To our knowledge, such a comprehensive evaluation has not been conducted so far in the field of multilingual ontology matching.

The rest of the paper is organised as follows. In Sect. 2, we first introduce the OntoFarm dataset and then we present its multilingual counterpart. We shortly discuss the hardness of MultiFarm and present the results that have been gathered in previous OAEI campaigns on OntoFarm. In Sect. 3, we present the evaluation setting used to carry out our experiments and list the tools we have evaluated. In Sect. 4, we finally describe the results of our experiments. We mainly focus on highly aggregated results due to the enormous amount of generated data. In Sect. 5, we conclude the paper and discuss directions for future work.

2 Background on MultiFarm

The MultiFarm dataset has been thoroughly described in [10]. It is available at http://web.informatik.uni-mannheim.de/multifarm/. The dataset is the multilingual version of the OntoFarm dataset [16], which has been used in previous OAEI campaigns in the Conference track. In the following, we shortly describe the OntoFarm dataset, explain how MultiFarm has been constructed, and roughly report about evaluation results of the OAEI Conference track.

2.1 OntoFarm

The OntoFarm dataset is based on a set of 16 ontologies from conference organisation domain. All contained ontologies differ in numbers of classes, properties, and in their DL expressivity. They are very suitable for ontology matching tasks since they were independently designed by different people who used various kinds of resources for ontology design:

- actual conferences and their web pages,
- actual software tools for conference organisation support, and
- experience of people with personal participation in organisation of actual conferences

Thus, the OntoFarm dataset describes a quite realistic matching scenario and has been successfully applied in the OAEI within the Conference track since 2006. In 2008, a first version of the reference alignments was created and then annually enriched and updated up to current 21 reference alignments built between seven (out of 16) ontologies. Each of them has between four to 25 correspondences. The relatively small number of correspondences in the reference alignments is based on the fact that the reference alignments contain only simple equivalence correspondences. Due to different modeling styles of the ontologies, for many concepts and properties thus no equivalent counterparts exist. This makes the matching task harder, however, it is also a typical characteristics of other matching scenarios.

2.2 MultiFarm

For generating the MultiFarm dataset, those seven OntoFarm ontologies, for which reference alignments are available, were manually translated into eight different languages (Chinese, Czech, Dutch, French, German, Portuguese, Russian, and Spanish). Since native speakers with a certain knowledge about the ontologies translated them, we do not expect any serious errors but of course they can never be excluded at all. Based on these translations, it is possible to re-create cross-lingual variants of the original test cases from the OntoFarm dataset as well as to exploit the translations more directly. Thus, the MultiFarm dataset contains two types of cross-lingual reference alignments.

We have depicted a small subset of the dataset shown in Fig. 1. This figure indicates the cross-lingual reference alignments between different ontologies, derived from original alignments and translations (type (i)), and cross-lingual reference alignments between the same ontologies, which are directly based on the translations or on exploiting transitivity of translations (type (ii)). Reference alignments of type (i) cover only a small subset of all concepts and properties. We have explained this above for the original test cases of the OntoFarm dataset. In contrast, for test cases of type (ii) there are (translated) counterparts for each concept and property.

Overall, the MultiFarm dataset has 36 × 49 test cases. 36 is a number of pairs of languages – each English ontology has its 8 language variants. 49 is the

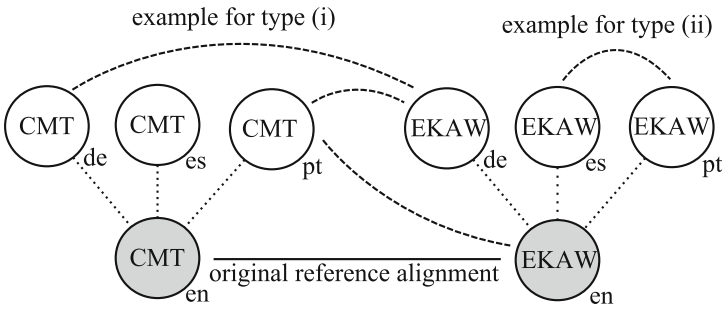

Fig. 1. Constructing MultiFarm from OntoFarm. Small subset that covers two ontologies and three translations. The solid line refers to a reference alignment of the OntoFarm dataset; dotted lines refer to translations; dashed lines refer to new cross-lingual reference alignments.

number of all reference alignments for each language pair. This is implied from the number of original reference alignments (21) which is doubled (42) due to the fact that there is a difference between CMT_{en}-$EKAW_{de}$ and CMT_{de}-$EKAW_{en}$ in comparison with the original test cases where the test cases CMT-$EKAW$ and $EKAW$-CMT are not distinguished. Additionally, we can also construct new reference alignments for matching each ontology on its translation which gives us seven additional reference alignments for each pair.

The main motivation for creating the MultiFarm dataset has been the ability to create a comprehensive set of test cases of type (i). We have especially argued in [10] that type (ii) test cases are not well suited for evaluating multilingual ontology matching systems, because they can be solved with very specific methods that are not related to the multilingual matching task.

2.3 Test Hardness

The OntoFarm dataset has a very heterogeneous character due to different modeling styles by various people. This leads to a high difficulty of the resulting test cases. For example, the object property `writtenBy` occurs in several OntoFarm ontologies. When only considering the labels, one would expect that a correspondence like `writtenBy` = `writtenBy` correctly describes that these object properties are equivalent. However, in ontology $\mathcal{O}_1$ the property indicates that a paper (domain) is written by an author (range), while in $\mathcal{O}_2$ the property describes that a review (domain) is written by a reviewer (range). Therefore, this correspondence is not contained in the reference alignment between $\mathcal{O}_1$ and $\mathcal{O}_2$. Similarly, comparing the English against the Spanish variant, there are the object properties `writtenBy` and `escrito por`. Pure translation would, similarly to the monolingual example, not result in detecting a correct correspondence. For that reason, the MultiFarm type (i) test cases go far beyond being a simple translation task.

The cross-lingual test cases of MultiFarm are probably much harder than the monolingual test cases of OntoFarm. Hence, it is important to know how matching systems perform on OntoFarm. These results can be understood as an upper bound that will be hard to top by results achieved for MultiFarm. In Fig. 2, we have depicted some results of previous OAEI campaigns in a precision/recall triangular graph. This graph shows precision, recall, and F-measure in a single plot. It includes the best (squares) and average (circles) results of the 2009, 2010, 2011 and 2011.5 Conference track as well as results of the three best ontology matching systems (triangles) from 2011.5. Best results are considered according to the highest F-measure which corresponds to exactly one ontology matching system for each year. In 2011, YAM++ achieved the highest F-measure that is why its cross sign overlaps with the light grey square depicting the best result of 2011.5. This matching system overcame 0.70 F-measure as a first system.

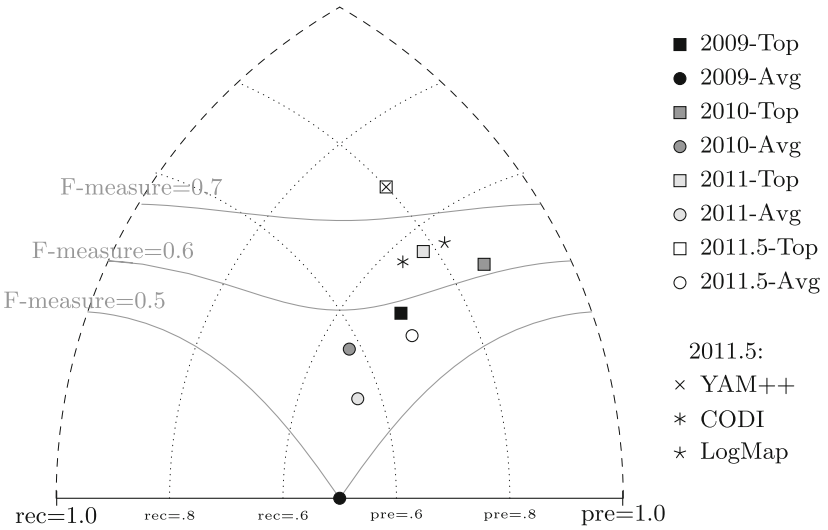

Fig. 2. Precision/recall triangular graph for the last four Conference tracks. Horizontal line depicts level of precision/recall while values of F-measure are depicted by areas bordered by corresponding lines F-measure = 0.[5|6|7].

On the one hand, Fig. 2 shows that there is an improvement every year, except the average results of 2011. Furthermore, average results of 2011.5 is almost the same as the results of the top matching system in 2009. A reason might be the availability of the complete dataset over several years. Since MultiFarm has not been used in the past, we expect that evaluation results also improve over the years. On the other hand, we can see that recall is not very high (0.63 in 2010, 0.60 in 2011 and 0.69 in 2011.5 for the best matching systems). This indicates that test cases of the OntoFarm dataset are especially difficult regarding recall measure.

3 Evaluation Settings

In the following, we explain how we executed our evaluation experiments and list the matching systems that have been evaluated.

3.1 Evaluation Workflow

Following a general definition, *matching* is the process that determines an *alignment* $\mathcal{A}$ for a pair of ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$. Besides the ontologies, there are other input parameters that are relevant for the matching process, namely: (*i*) the use of an input alignment $\mathcal{A}'$, which is to be extended or completed by the process; (*ii*) parameters that affect the matching process, for instance, weights and thresholds; and (*iii*) external resources used by the matching process, for instance, common knowledge and domain specific thesauri.

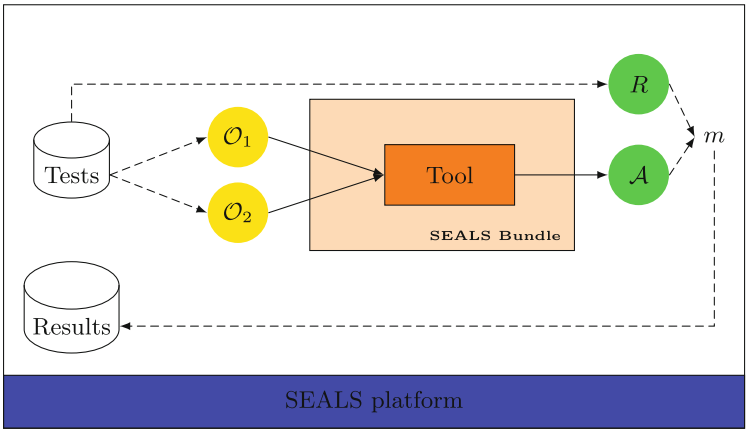

Fig. 3. Execution of tools.

In this paper, we focus on evaluating a standard matching task. (*i*) In most of our experiments, we do not modify the parameters that affect the matching process. For two systems, we made an exception from this rule and report very briefly on the results. (*ii*) We do not use an additional input alignment at all. Note that most systems do not support such a functionaility. (*iii*) We put no restriction on the external resources that are taken into account by the evaluated systems. Thus, we use the system standard settings for our evaluation. However, we obviously focus on the matching process where labels and annotations of $\mathcal{O}_1$ and $\mathcal{O}_2$ are described in different languages.

The most common way to evaluate the quality of a matching process is to evaluate the correctness (precision) and completeness (recall) of its outcome $\mathcal{A}$ by comparing $\mathcal{A}$ against a reference alignment $\mathcal{R}$. Since 2010, in the context of OAEI campaigns, the process of evaluating matching systems has been automated thanks to the SEALS platform (Fig. 3). For OAEI 2011 and OAEI 2011.5,

participants have been invited to wrap their tools into a format that can be executed by the platform, i.e. the matching process is not conducted by the tool developer but by the organisers of an evaluation using the platform. For the purpose of this paper, we benefit from the large number of matching tools that become available for our evaluation. Furthermore, evaluation test cases are available in the SEALS repositories and can be used by everyone. Thus, all of our experiments can be completely reproduced.

3.2 Evaluated Matching Systems

As stated before, a large set of matching systems has already been uploaded to the platform in the context of OAEI 2011. We apply most of these tools to the MultiFarm dataset. In particular, we evaluated the tools AROMA [2], CIDER [5], CODI [6], CSA [15], LogMap and LogMapLt [7], MaasMatch [13], MapSSS [1], YAM++ [11] and Lily [17]. For most of these tools, we used the version submitted to OAEI 2011. However, some tool developers have already submitted a new version with some modifications between OAEI 2011 and OAEI 2011.5. This is the case for CODI, LogMap and MapSSS. Moreover, the developer of LogMap has additionally uploaded a lite version of their matching systems called LogMapLt.

We also included three OAEI 2011.5 participants: WeSeE and AUTOMSv2 [9] and the OAEI 2011.5 version of YAM++. WeSeE and YAM++ use Microsoft Bing to translate labels contained in the input ontologies to English. The translated English ontologies are then matched using standard matching procedures of WeSeE and YAM++. AUTOMSv2 re-uses a free Java API named WebTranslator to translate the ontologies to English. This process is performed before AUTOMSv2 profiling, configuration and matching methods are executed, so their input will consider only English-labeled copies of ontologies.

Since MultiFarm is based on the Conference dataset, we provide an overview table regarding performance of evaluated matching systems within last three editions of the Conference track, see Table 1. The last column (ML) of the table indicates whether systems uses multilingual components in the matching process. There have also been some systems participating in OAEI 2011 and OAEI 2011.5 that are not listed here. We have not added them to the evaluation for different reasons. Some of these systems cannot finish the MultiFarm matching process in less than several weeks while others generate empty alignments for nearly all matching tasks or terminate with an error. With respect to the OAEI 2011.5 participants, we have only added those systems that use mulitlingual techniques.

We have already explained that the MultiFarm data set is a comprehensive collection of testcases. For that reason we executed some of the tools in paralell on top of the SEALS platform. While systems as LogMap finished the MultiFarm dataset in less than 30 min, other systems required up to several days. However, reporting runtimes is beyond the scope of this paper.

Table 1. Performance of evaluated matching systems within last three editions of the Conference track (P = precision, R = recall, F = f-measure).

Matcher	OAEI 2010			OAEI 2011			OAEI 2011.5			ML
	P	F	R	P	F	R	P	F	R	
AROMA	.36	.42	.49	.35	.40	.46	*N/A*			
CIDER	*N/A*			.64	.53	.45	*N/A*			
CODI	.86	.62	.48	.74	.64	.57	.74	.64	.57	
CSA	*N/A*			.50	.55	.60	*N/A*			
LogMap	*N/A*			.84	.63	.50	.82	.66	.55	
LogMapLt	*N/A*			*N/A*			.73	.59	.50	
MaasMatch	*N/A*			.83	.56	.42	.74	.54	.42	
MapSSS	*N/A*			.55	.51	.47	.50	.50	.51	
YAM++	*N/A*			.78	.65	.56	.80	.74	.69	√
Lily	*N/A*			.36	.41	.47	*N/A*			
AUTOMSv2	*N/A*			*N/A*			.75	.52	.40	√
WeSeE	*N/A*			*N/A*			.67	.55	.46	√

4 Results

In the following, we discuss the results on different perspectives. First, we aggregate the results obtained for all pairs of test cases (and languages) in Sect. 4.1. Then we focus on different pairs of languages in Sect. 4.2. In both sections we report only on results for those systems that are not specifically designed for multilingual matching. In Sect. 4.3, we finally evaluate those three OAEI 2011.5 systems that use specific multilingual components.

4.1 Differences in Test Cases

As explained in Sect. 2, the dataset can be divided in (i) those test cases where the ontologies to be matched are translations of different ontologies and (ii) those test cases where the same original ontology has been translated into two different languages and the translated ontologies have to be matched. We display the results for test cases of type (i) on the left and those for type (ii) on the right of Table 2. We have ordered the systems according to the F-measure for the test cases of type (i). The best results, in terms of F-measure, are achieved by CIDER (18 %) followed by CODI (13 %), LogMap (11 %) and MapSSS (10 %). CIDER has both better precision and recall scores than any other system. Compared to the top-results that have been reported for the original Conference dataset (F-measure > 60 %), the test cases of the MultiFarm dataset are obviously much harder. However, an F-measure of 18 % is already a remarkable result given the fact that we executed CIDER in its default setting.

Table 2. Results aggregated per matching system.

Matcher	(i) Different ontologies				(ii) Same ontologies			
	Size	P	F	R	Size	P	F	R
CIDER	1433	.12	.18	.42	1090	.66	.12	.06
CODI	923	.08	.13	.43	7056	.77	.59*	.48
LogMap	826	.39	.11	.06	469	.71	.06	.03
MapSSS	2513	.16	.10	.08	6008	.97	.67*	.51
LogMapLt	826	.26	.07	.04	387	.56	.04	.02
MaasMatch	558	.24	.05	.03	290	.56	.03	.01
CSA	17923	.02	.03	.06	8348	.49	.42*	.36
YAM++$_{2011}$	7050	.02	.03	.03	4779	.22	.13*	.09
Aroma-	0	-	-	.00	207	.54	.02	.01
Lily	0	-	-	.00	11	1.00	.00	.00

The outcomes for test cases of type (ii) differ significantly. In particular, the results of MapSSS (67 % F-measure) are surprisingly compared to the results presented for test cases of type (i). This system can leverage the specifics of type (ii) test cases to cope with the problem of matching labels expressed in different languages. Similar to MapSSS, we also observe a higher F-measure for CODI, CSA, and YAM++. We have marked those systems with an asterisk. Note that all these systems have an F-measure of at least five times higher than the F-measure for test cases of type (i). For all other systems, we observe a slightly decreased F-measure comparing test cases of type (i) with type (ii).

Again, we have to highlight the differences between both types of test cases. Reference alignments of type (i) cover only a small fraction of all concepts and properties described in the ontologies. This is not the case for test cases of type (ii). Here, we have complete alignments that connect each concept and property with an equivalent counterpart in the other ontology. There seems to be a clear distinction between systems that are configured to generate complete alignments in the absence of (easy) usable label description, and other systems that focus on generating good results for test cases of type (i).

Comparing these results with the results for the OAEI 2011 Benchmark track, it turns out that all systems marked with an asterisk have been among the top five systems of this track. All Benchmark test cases have a similar property, namely, their reference alignments contain for each entity of the smaller ontology exactly one counterpart in the larger ontology. An explanation for this can be that these systems have been developed or at least configured to score well for the Benchmark track. For that reason, they generate good results for test cases of type (ii), while their results for test cases of type (i) are less good. MapSSS and CODI are an exception. These systems generate good results for both test cases of type (i) and (ii).

4.2 Differences in Languages

Besides aggregating the results per matcher, we have analysed the results per pair of languages (Table 3), for the case where different ontologies are matched (type (i) in Table 2). We have also compared the matchers with a simple edit distance algorithm on labels (edna).

Table 3. Results (F-measure) per pairs of languages for different ontologies.

pairs	edna	Aroma	CIDER	CODI	CSA	Lily	LogMap	LogLt	MaasMatch	MapSSS	YAM++$_{2011}$	average
cz-de	.01	-	.12	.11	.03	-	.09	.09	.02	.07	.03	.06
cz-en	.01	-	.20	.12	.04	-	.06	.04	.03	.08	-	.07
cz-es	.00	-	.14	.13	.02	-	.11	.11	-	.11	.03	.08
cz-fr	.01	-	.08	.01	.01	-	.01	.01	.01	.01	.03	.02
cz-nl	.00	-	.09	.09	.04	-	.04	.04	.04	.05	.03	.05
cz-pt	.00	-	.15	.15	.04	-	.13	.13	.02	.12	.04	.09
de-en	.01	-	.31	.22	.03	-	.22	.20	.20	.16	-	**.17**
de-es	.01	-	.25	.20	.02	-	.19	.06	-	.15	.03	**.11**
de-fr	.00	-	.18	.18	.01	-	.17	.04	.04	.13	.03	.09
de-nl	.01	-	.22	.08	.03	-	.05	.04	.04	.15	.03	.07
de-pt	.01	-	.10	.09	.03	-	.07	.07	.01	.06	.04	.05
en-es	.00	-	.25	.24	.03	-	.18	.04	.04	.18	-	**.12**
en-fr	.01	-	.20	.24	.03	-	.19	.04	.04	.13	-	**.11**
en-nl	.01	-	.22	.10	.04	-	.07	.10	.07	.15	-	**.10**
en-pt	.00	-	.15	.11	.06	-	.06	.06	.06	.07	-	.07
es-fr	.01	-	.29	.07	.02	-	.06	.01	.04	.06	.03	.07
es-nl	.01	-	.07	.01	.02	-	-	-	-	.01	.02	.02
es-pt	.01	-	.29	.26	.06	-	.27	.23	.09	.23	.03	**.16**
fr-nl	.01	-	.23	.14	.02	-	.13	.12	.13	.11	.03	**.10**
fr-pt	.00	-	.11	.06	.02	-	.06	-	.04	.02	.03	.04
nl-pt	.00	-	.02	.04	.03	-	.01	.01	.02	.02	.04	.02
average	.01		.17	.13	.03		.11	.08	.05	.10	.03	.08

With exception of Aroma and Lily, which are not able to deal with the complexity of the matching task, for most of the test cases no matcher has lower F-measure than edna. For some of them, however, LogMap, LogMapLt, MaasMatch and YAM++, respectively, have not provided any alignment. YAM++ has a specific behaviour and is not able to match the English ontologies to any other languages. For the other matchers, it (incidentally) happens mostly for the pairs of languages that do not share the same root language (e.g. es-nl or de-es). The exception is LogMapLt, which is not able to identify any correspondence between fr-pt, even if these languages have the same root language (e.g. Latin)

and thus have a similar vocabulary. It could be expected that matchers should be able to find a higher number of correspondences for the pairs of languages where there is an overlap in their vocabularies because most of the matcher apply some label similarity strategy. However, it is not exactly the case in MultiFarm. The dataset contains many complex correspondences that cannot be found by a single translation process or by string comparison. This can be partially corroborated by the very low performance of edna in all test cases.

Looking at the results for each pair of languages, per matcher, the best five F-measures are obtained for de-en (31 %), es-fr/es-pt (29 %), de-es/en-es (25 %), all for CIDER, en-es/en-fr (24 %), for CODI, and fr-nl (23 %) again for CIDER. We could observe that 3 ahead pairs contain languages with some degree of overlap in their vocabularies (i.e., de-en, es-fr, es-pt). For each individual matcher, seven out of eight matchers have their best scores for these pairs (exception is YAM++ that scores better for cz-pt and de-pt), with worst scores in cz-fr, es-nl, which have very different vocabularies.

When aggregating the results per pair of languages, that order is mostly preserved (highly affected by CIDER): de-en (17 %), es-pt (16 %), en-es (12 %), de-es/en-fr (11 %), followed by fr-nl/en-nl (10 %). The exception is for the pair es-fr, where the aggregated F-measure decreases to 7 %. Again, the worst scores are obtained for cz-fr, nl-pt and es-nl. We can observe that, for most of the cases, the features of the languages (i.e., their overlapping vocabularies) have an impact in the matching results. However, there is no universal pattern and we have cases with similar languages where systems score very low (fr-pt, for instance). This has to be further analysed looking at the individual pairs of ontologies.

4.3 Translation Based Techniques

We have also analysed three matching systems (YAM++, AUTOMSv2 and WeSeE), participating in OAEI 2011.5, which first translate both source ontologies into English. The results, per pair of languages where different ontologies are matched (type (i) in Table 2), are reported in Table 4. These three matching systems clearly outperform all the other systems which do not use any specific method to deal with multilingual ontologies (cf. Table 3). Looking at the average results, the best results are achieved by YAM++ (0.41) followed by AUTOMSv2 (0.36) and WeSeE (0.27). However while YAM++ and WeSeE managed to match all eight different languages in the MultiFarm dataset, AUTOMSv2 did not manage to match ontologies in Chinese, Czech and Russian languages.

Looking at the results for each pair of languages, per matcher, the best five F-measures are obtained for en-fr (61 %), cz-en (58 %), cz-fr (57 %), en-pt (56 %), and en-nl/cz-pt/fr-pt (55 %). We can observe a positive effect of the translation step, since (besides the differences in the structure of the source ontologies) most of these language pairs do not have overlapping vocabularies (cz-pt or cz-fr, for instance). Furthermore, results with the same translator can differ depending on further matching system's components as demonstrated by YAM++ and WeSeE. YAM++ outperforms WeSeE for most of the pairs. When looking at the average of these three systems, we have the following pairs ranking: en-fr (47 %), en-pt

Table 4. Performance of the three OAEI 2011.5 systems that implemented specific multilingual methods.

Matcher/Pair	YAM++			AUTOMSv2			WeSeE			Average		
	P	F	R	P	F	R	P	F	R	P	F	R
cn-cz	.44	.32	.25	-	-	-	.14	.18	.24	.29	.25	.24
cn-de	.4	.32	.27	-	-	-	.11	.15	.21	.26	.23	.24
cn-en	.47	.38	.32	-	-	-	.41	.29	.22	.44	.33	.27
cn-es	.58	.2	.12	-	-	-	.1	.13	.2	.34	.17	.16
cn-fr	.43	.35	.3	-	-	-	.14	.17	.23	.28	.26	.26
cn-nl	.42	.35	.29	-	-	-	.09	.11	.17	.25	.23	.23
cn-pt	.34	.27	.22	-	-	-	.09	.12	.18	.22	.2	.2
cn-ru	.4	.3	.24	-	-	-	.09	.12	.17	.24	.21	.21
cz-de	.51	.47	.45	-	-	-	.16	.21	.3	.33	.34	.37
cz-en	.6	.58	.57	-	-	-	.51	.45	.4	.55	.51	.48
cz-es	.58	.2	.12	-	-	-	.21	.28	.43	.39	.24	.28
cz-fr	.58	.57	.57	-	-	-	.2	.26	.37	.39	.42	.47
cz-nl	.54	.53	.53	-	-	-	.16	.21	.3	.35	.37	.41
cz-pt	.54	.55	.56	-	-	-	.18	.24	.34	.36	.39	.45
cz-ru	.55	.55	.55	-	-	-	.2	.25	.35	.37	.4	.45
de-en	.55	.47	.42	.8	.39	.26	.52	.44	.37	.62	.43	.35
de-es	.58	.2	.12	.73	.37	.24	.16	.21	.33	.49	.26	.23
de-fr	.5	.46	.43	.86	.32	.2	.19	.24	.33	.51	.34	.32
de-nl	.5	.45	.41	.78	.39	.26	.18	.24	.35	.48	.36	.34
de-pt	.45	.39	.34	.82	.35	.22	.2	.26	.37	.49	.33	.31
de-ru	.56	.51	.47	-	-	-	.15	.19	.26	.36	.35	.37
en-es	.57	.21	.13	.61	.42	.32	.56	.5	.46	.58	.38	.3
en-fr	.6	.61	.62	.54	.3	.21	.58	.51	.45	.57	.47	.43
en-nl	.57	.55	.52	.6	.34	.24	.53	.45	.4	.57	.45	.38
en-pt	.58	.56	.54	.61	.37	.27	.53	.47	.43	.57	.47	.41
en-ru	.6	.55	.5	-	-	-	.58	.49	.43	.59	.52	.46
es-fr	.54	.2	.12	.62	.37	.26	.27	.35	.5	.47	.3	.3
es-nl	.48	.16	.1	.49	.35	.27	.16	.22	.37	.38	.24	.24
es-pt	.61	.25	.15	.54	.44	.37	.22	.3	.47	.46	.33	.33
es-ru	.55	.21	.13	-	-	-	.15	.21	.33	.35	.21	.23
fr-nl	.54	.53	.51	.54	.27	.18	.2	.26	.38	.43	.35	.36
fr-pt	.55	.55	.55	.63	.35	.24	.24	.31	.44	.47	.4	.41
fr-ru	.56	.52	.5	-	-	-	.19	.24	.34	.37	.38	.42
nl-pt	.52	.49	.46	.61	.38	.28	.17	.23	.35	.43	.36	.36
nl-ru	.56	.52	.5	-	-	-	.16	.22	.33	.36	.37	.41
pt-ru	.54	.53	.52	-	-	-	.16	.21	.3	.35	.37	.41
Average	**.52**	**.41**	**.37**	**.65**	**.36**	**.25**	**.25**	**.27**	**.34**	**.42**	**.34**	**.33**

(47 %), en-nl (45 %), de-en (43 %) and fr-pt (40 %). We can observe that for these pairs, the English language is in most of the pairs. It is somehow expected because these matchers use English as the pivot language in the translation process and the pure translation results are less penalised with regards to the lack of complementary strategies, such as translation disambiguation.

5 Discussion

Some of the reported results are relevant for multilingual ontology matching in general, while others help us to understand the characteristics of the MultiFarm dataset. The latter ones are relevant for any further evaluation that builds on the dataset. Moreover, we can also draw some conclusions that might be important for the use of datasets in the general context of ontology matching evaluation.

Exploiting structural information. Very good results for test cases of type (ii) can be achieved by methods non-specific to multilingual ontology matching. The result of MapSSS is an interesting example. This was also one of the main reasons why the MultiFarm dataset has been constructed as a comprehensive collection for test cases of type (i) and (ii). We suggest to put a stronger focus on test cases of type (i) in the context of evaluating multilingual ontology matching techniques. Otherwise, it remains unclear whether the measured results are based on multilingual techniques or on exploiting that the matched ontologies can interpreted as versions of the same ontology.

Finding a good configuration. Our results show that state-of-the-art matching systems are not very well suited for the tasks of matching ontologies described in different languages, especially when executed in their default setting. We started another set of experiments by running some tools (CODI, LogMap, Lily) in a manually configured setting better suited for the matching task. A first glimpse, the results shows that it is possible to increase the average F-measure up to a value of 26 %. Thus, we are planning to further investigate the influence of configurations on multilingual matching tasks within more extensive experiments.[4]

The role of language features. We cannot neglect certain language features (like their overlapping vocabularies) in the matching process. Once most of the matchers take advantage of label similarities it is likely that it may be harder to find correspondences between Czech and Portuguese ontologies than Spanish and Portuguese ones. In our evaluation, for most of the systems, the better performance where incidentally observed for the pairs of languages that have some degree of overlap in their vocabularies. This is somehow expected, however, we could find exceptions to this behavior. In fact, MultiFarm requires systems exploiting more sophisticated matching strategies than label similarity and

[4] We would like to thank Ernesto Jimenez-Ruiz (LogMap [7]) and Peng Wang (Lily [17]) for supporting us with a quick hint about a good, manually modified configuration for running their systems on MultiFarm.

for many ontologies in MultiFarm it is the case. To some extent we exploited automatic translation strategies by evaluating results from systems exploiting translations of ontologies into English language.

Test Difficulty. We can give the following simplified conclusion related to test difficulty. For the conference track top systems generate results with an average F-measure of 0.6 to 0.7 with better precision and worse recall. State of the art matching systems, without multilingual component, generate in their default setting in average an F-measure between 0 and 0.2 for the MultiFarm testcases. Using a well-chosen configuration, this value can increase up to 0.25 (based on very low recall values). A system that uses a preceding translation step, can achieve an F-measure between 0.3 and 0.4. These results are still based on slightly higher precision scores, however, the differences between precision and recall are less significant.

Implications on analyzing OAEI results. Aside from the topic of multilingual ontology matching, the results implicitly emphasise the different characteristics of test cases of type (i) and (ii). An example can be found when comparing results for the OAEI Benchmark and Conference track. The Benchmark track is about matching different versions (some slightly modified, some heavily modified) of the same ontology. The Conference dataset is about matching different ontologies describing the same domain. This difference finds its counterparts in the distinction between type (i) and type (ii) ontologies in the MultiFarm dataset. Without taking this distinction into account, it is not easy to draw valid conclusions on the generality of measured results.

6 Future Work

Even though we reported about diverse aspects, we could not analyse or evaluate all interesting issues. The following listing shows possible extensions and improvements for further evaluations based on MultiFarm:

- Executing matching systems with a specifically tailored configuration;
- Exploiting other approaches than pure translation strategies (disambiguation, use of multilingual lexicons, multilingual comparable corpora) and evaluate their impact on the matching process;
- Analysing the role of diacritics: in some languages, the same word written with or without accent can have a different meaning, e.g., in French ‘où’ (where) is different from ‘ou’ (or);
- Exploiting ontology population strategies for creating MultiFarm instances and take advantage of instance-level matching approaches; and evaluate how these approaches can help in the multilingual matching process.

Although we have many different ways to improve the multilingual matching task, we have shown that the MultiFarm dataset is a useful, comprehensive, and a difficult dataset for evaluating ontology matching systems. We strongly

recommend to apply this resource and to compare measured results against the results presented in this paper. In particular, we encourage developers of ontology matching systems, specifically designed to match ontologies described in different languages, to make use of the dataset and to report about achieved results.

Acknowledgements. Some of the authors are partially supported by the SEALS project (IST-2009-238975). Ondřej Šváb-Zamazal has been partially supported by the CSF grant no. P202/10/0761.

References

1. Cheatham, M.: MapSSS results for OAEI 2011. In: Proceedings of the 6th ISWC Workshop on Ontology Matching (OM), pp. 184–189 (2011)
2. David, J.: Aroma results for OAEI 2009. In: Proceedings of the 4th ISWC Workshop on Ontology Matching (OM), pp. 147–152 (2009)
3. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology alignment evaluation initiative: six years of experience. In: Spaccapietra, S. (ed.) Journal on Data Semantics XV. LNCS, vol. 6720, pp. 158–192. Springer, Heidelberg (2011)
4. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
5. Gracia, J., Bernad, J., Mena, E.: Ontology matching with CIDER: evaluation report for OAEI 2011. In: Proceedings of the 6th ISWC Workshop on Ontology Matching (OM), pp. 126–133 (2011)
6. Huber, J., Sztyler, T., Noessner, J., Meilicke, C.: Codi: Combinatorial optimization for data integration: results for OAEI 2011. In: Proceedings of the 6th ISWC Workshop on Ontology Matching (OM), pp. 134–141 (2011)
7. Jimenez-Ruiz, E., Morant, A., Grau, B.C.: LogMap results for OAEI 2011. In: Proceedings of the 6th ISWC Workshop on Ontology Matching (OM), pp. 163–170 (2011)
8. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. Knowl. Eng. Rev. **18**(1), 1–31 (2003)
9. Kotis, K., Katasonov, A., Leino, J.: Aligning smart and control entities in the IoT. In: Andreev, S., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART 2012. LNCS, vol. 7469, pp. 39–50. Springer, Heidelberg (2012)
10. Meilicke, C., Castro, R.G., Freitas, F., van Hage, W.R., Montiel-Ponsoda, E., de Azevedo, R.R., Stuckenschmidt, H., Šváb-Zamazal, O., Svátek, V., Tamilin, A., Trojahn, C., Wang, S.: Multifarm: A benchmark for multilingual ontology matching. J. Web Semant. **2**(1), 3–10 (2011)
11. Ngo, D., Bellasene, Z., Coletta, R.: YAM++ - results for OAEI 2011. In: Proceedings of the 6th ISWC Workshop on Ontology Matching (OM), pp. 228–235 (2011)
12. Rahm, E., Bernstein, P.: A survey of approaches to automatic schema matching. VLDB J. **10**(4), 334–350 (2001)
13. Schadd, F., Roos, N.: Maasmatch results for OAEI 2011. In: Proceedings of the 6th ISWC Workshop on Ontology Matching (OM), pp. 171–178 (2011)

14. Spohr, D., Hollink, L., Cimiano, P.: A machine learning approach to multilingual and cross-lingual ontology matching. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 665–680. Springer, Heidelberg (2011)
15. Tran, Q.-V., Ichise, R., Ho, B.-Q.: Cluster-based similarity aggregation for ontology matching. In: Proceedings of the 6th ISWC Workshop on Ontology Matching (OM), pp. 142–147 (2011)
16. Šváb, O., Svátek, V., Berka, P., Rak, D., Tomášek, P.: Ontofarm: Towards an experimental collection of parallel ontologies. In: Poster Track of ISWC 2005 (2005)
17. Wang, P.: Lily results on SEALS platform for OAEI 2011. In: Proceedings of the 6th ISWC Workshop on Ontology Matching (OM), pp. 156–162 (2011)