A Practical Implementation of Semantic Precision and Recall

Daniel Fleischhacker and Heiner Stuckenschmidt
KR and KM Research Group
University of Mannheim
Mannheim, Germany
dfleisch@mail.uni-mannheim.de, heiner@informatik.uni-mannheim.de

***Abstract*—The systematic evaluation of ontology alignments still faces a number of problems. One is the argued inadequacy of traditional quality measures adopted from the field of information retrieval. In previous work, Euzenat and others have proposed notions of semantic precision and recall that are supposed to better reflect the true quality of an alignment by considering its deductive closure rather than the explicitly stated correspondences. So far, these measures have only been investigated in theory. In this paper, we present the first implementation of a restricted version of semantic precision and recall as well as experiments in using it, we conducted on the results of the 2008 OAEI campaign.**

***Index Terms*—Semantic Precision and Recall, Alignment Evaluation**

I. Motivation

Evaluating the quality of ontology alignments[1] is an important topic. Despite several years of effort in benchmarking ontology matching tools (e.g. [1]), there are still some open problems. In particular, standard measures of precision and recall that have been adopted from the information retrieval community have some weaknesses in this context. First of all, these measures require the existence of a complete reference mapping, the mapping to be evaluated can be compared with. This is an inherent problem of precision and recall and can also be found in other areas including traditional information retrieval. In order to avoid this problem, several researchers have proposed alternative evaluation measures and approaches [2], [3]. These alternative approaches, however, cannot completely replace a meaningful evaluation of precision and recall, because they either only provide qualitative judgements or upper bounds for the quality of a mapping.

The direct application of precision and recall to ontology mappings comes with some problems related to the logical nature of ontologies. In particular, it is often possible to derive implicit mappings from given ones, especially, if the mappings considered are not limited to equivalence relations between classes and relations but for instance also include subsumption mappings. In order to overcome this problem, Euzenat and others [4], [5] have proposed semantic precision and recall measures that given two mappings do not only compare these on syntactic level but also take derivable mappings into account. So far, these measures have only been investigated theoretically.

The goal of the work reported in this paper is to put semantic precision and recall into practice. For this purpose, we propose simplified versions of semantic precision and recall. We implemented these simplified version of semantic precision and recall and tested it on results of the Ontology Alignment Evaluation Initiative. Comparing the results with traditional precision and recall reveals that using the semantic version of precision and recall actually makes a difference not only with respect to the absolute values for precision and recall but also with respect to the relative quality of a matching system compared to others, providing new insight in the actual quality of matching systems.

The paper is structured as follows. After introducing some basic definitions in the next section we define semantic precision and recall in section 3. Our experiments with computing semantic versions of precision and recall and comparing it to the traditional ones are reported in section 4. We conclude with a discussion of related work in section 5 and a summary of our results and future work.

II. Preliminaries

In this section we introduce the terms required to define the semantically founded measures as used in the remaining paper. Foremost, we give a more formal definition for the semantics of alignments as used in [4].

Definition 1 (Correspondence, Alignment): Given two ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$. A correspondence c between two entities $e_1 \in \mathcal{O}_1$ and $e_2 \in \mathcal{O}_2$ is a four-tuple $c = \langle e_1, e_2, r, m \rangle$ where r is the relation holding between these entities and m is a value describing the confidence in the validity of this correspondence.

An alignment A for ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ is a set of correspondences between entities from these two ontologies.

Assuming that a model theoretic semantics for the ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ is given, we can define the semantics of an alignment in terms of a translation into a logical theory as done by Meilicke and Stuckenschmidt [6]. The following definitions given in this section are also roughly equivalent to those introduced in their work.

Definition 2 (Semantics of Alignments): Given two ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ and an alignment A between these two

[1]Throughout this paper, we use the words "alignment" and "mapping" interchangeably.

ontologies. Let L be a suitable logical language and S(L) the set of possible sentences of L, then a reductionistic alignment semantics is given by L and a pair of functions $\langle L, \mathrm{ext}, \mathrm{trans}\rangle$. ext is an extension function such that $\mathrm{ext}(\mathcal{O}_1, \mathcal{O}_2) \subset S(L)$ and trans a translation function such that $\mathrm{trans}(A) \subset S(L)$.

In the following, we always assume that L is the description logic $\mathcal{SHOIQ}(D)$ which is the basis for the Web Ontology Language OWL [7] and omit the language L in all definitions.

Having this definition the next step requires the possibility to interpret alignments together with ontologies. This is provided by the following notion of an aligned ontology.

Definition 3 (Aligned Ontology): Given two ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ and an alignment A, their aligned ontology for a specific reductionistic semantics $S = \langle \mathrm{ext}, \mathrm{trans}\rangle$ is given by

$$\mathrm{ext}(\mathcal{O}_1, \mathcal{O}_2) \cup \mathrm{trans}(A)$$

We refer to this aligned ontology with the ontologies, alignment and semantics given above as $\mathcal{O}_1 \cup_A^S \mathcal{O}_2$.

Having determined the logical language, we can now define two concrete variants of alignment semantics. At first, we present the most natural semantics which simply translates the correspondences directly into the ontology language.

Definition 4 (Natural Semantics): Given ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ and an alignment A between these ontologies. The natural semantics is defined as

$$S_n = \langle \mathrm{ext}_n, \mathrm{trans}_n\rangle$$

where $\mathrm{ext}_n(\mathcal{O}_1, \mathcal{O}_2) \longmapsto \mathcal{O}_1 \cup \mathcal{O}_2$ and $\mathrm{trans}_n(A) \longmapsto \{t_n(c) \,|\, c \in A\}$ is given by

$$t_n(c) \longmapsto \begin{cases} e_1 \sqsubseteq e_2 & \text{if } c = \langle e_1, e_2, \leq\rangle \\ e_1 \equiv e_2 & \text{if } c = \langle e_1, e_2, =\rangle \\ e_1 \sqsupseteq e_2 & \text{if } c = \langle e_1, e_2, \geq\rangle \end{cases}$$

In practice there are many alignments which map dataproperties to objectproperties or vice versa. Due to the disjointness of both property types' domains in OWL-DL, the aligned ontology gained from such an alignment is incoherent. Unfortunately, it is not always possible to prevent such correspondences because they are useful and often required to resolve differences in the modeling of the different ontologies. To dampen the effect of such alignments and avert the incoherence of the aligned ontology, we use the variant of natural semantics introduced by Meilicke in Carracciolo et al. [1] and formalized by Meilicke and Stuckenschmidt [6]. We refer to this variant as pragmatic semantics.

Definition 5 (Pragmatic Semantics): Given ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ and a corresponding alignment A, the pragmatic semantics is defined as

$$S_p = \langle \mathrm{ext}_p, \mathrm{trans}_p\rangle$$

where $\mathrm{ext}_p(\mathcal{O}_1, \mathcal{O}_2) \longmapsto \mathcal{O}_1 \cup \mathcal{O}_2$ and $\mathrm{trans}_p(A) \longmapsto \{t_p(c) \,|\, c \in A\}$. For the translation function we define

$$t_p(c) \longmapsto \begin{cases} m_p(e_1) \sqsubseteq m_p(e_2) & \text{if } c = \langle e_1, e_2, \leq\rangle \\ m_p(e_1) \equiv m_p(e_2) & \text{if } c = \langle e_1, e_2, =\rangle \\ m_p(e_1) \sqsupseteq m_p(e_2) & \text{if } c = \langle e_1, e_2, \geq\rangle \end{cases}$$

and

$$m_p(e) \longmapsto \begin{cases} e & \text{if } e \text{ is a concept} \\ \exists e.\top & \text{otherwise} \end{cases}$$

The pragmatic semantics guarantees that only classes are mapped onto each other and thus it is not possible to cause incoherence due to correspondences across property domains.

III. Restricted Semantic Precision and Recall

With these notions of alignment semantics we are now able to define some more terms required to define the measures as used here. The following notions are based on the definition of semantic precision and recall by Euzenat [4] but are adapted to our different understanding of alignment semantics. The first concept to introduce formally is the concept of α-consequences.

Definition 6 (α-Consequence of an Alignment): Given ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$, a corresponding alignment A and a reductionistic semantics $S = \langle \mathrm{ext}, \mathrm{trans}\rangle$ a correspondence c is an α-consequence of $\mathcal{O}_1$, $\mathcal{O}_2$, S and A if and only if

$$\mathcal{O}_1 \cup_A^S \mathcal{O}_2 \vDash \mathrm{trans}(c)$$

In this case we say $A \vDash^S_{\mathcal{O}_1,\mathcal{O}_2} c$.

Applying this definition to complete alignments instead of single correspondences we get the closure of an alignment which resembles the sets of α-correspondences used by Euzenat.

Definition 7 (Closure of an Alignment): Given ontologies $\mathcal{O}_1$, $\mathcal{O}_2$ and a reductionistic semantics S the closure Cn of an alignment A is given by

$$\mathrm{Cn}^S_{\mathcal{O}_1,\mathcal{O}_2}(A) = \{c \,|\, A \vDash^S_{\mathcal{O}_1,\mathcal{O}_2} c\}$$

Furthermore, for an additional alignment B the statement $A \vDash^S_{\mathcal{O}_1,\mathcal{O}_2} B$ holds, iff

$$\mathrm{Cn}^S_{\mathcal{O}_1,\mathcal{O}_2}(A) \supseteq \mathrm{Cn}^S_{\mathcal{O}_1,\mathcal{O}_2}(B)$$

Obviously, the closure of an alignment is an alignment itself because it is a set of correspondences.

In order to avoid the problems with the original definition, we distinguish between complex and non-complex correspondences.

Definition 8 (Complex/Non-Complex Correspondences): A correspondence $c = \langle e_1, e_2, r\rangle$ is non-complex if both e_1 and e_2 are atomic entities (i.e. concept or relation names) from the corresponding ontologies. Otherwise c is complex.

An alignment is said to be non-complex if it contains only non-complex correspondences.

Now we introduce a restricted variant of the ideal semantic precision and recall measures as given by Euzenat [4].

Definition 9 (Restricted Semantic Precision and Recall): Given consistent ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$, two non-complex alignments between these two ontologies, namely the reference alignment R and the alignment A which is to be evaluated, and a reductionistic semantics S. Further, let the aligned ontologies $\mathcal{O}_1 \cup_R^S \mathcal{O}_2$ and $\mathcal{O}_1 \cup_A^S \mathcal{O}_2$ be consistent.

The restricted semantic precision is defined as

$$\mathrm{P}_r(A,R) = \frac{|\mathrm{Cn}^S_{\mathcal{O}_1,\mathcal{O}_2}(A) \cap \mathrm{Cn}^S_{\mathcal{O}_1,\mathcal{O}_2}(R)|}{|\mathrm{Cn}^S_{\mathcal{O}_1,\mathcal{O}_2}(A)|}$$

and the restricted semantic recall is defined as

$$\mathrm{R}_r(A,R) = \frac{|\mathrm{Cn}^S_{\mathcal{O}_1,\mathcal{O}_2}(A) \cap \mathrm{Cn}^S_{\mathcal{O}_1,\mathcal{O}_2}(R)|}{|\mathrm{Cn}^S_{\mathcal{O}_1,\mathcal{O}_2}(R)|}$$

These measures prevent most problems arising for the ideal semantic precision and recall measures by restricting the considered alignments to a subset of all possible alignments. Due to the restriction to non-complex alignments the closures of these alignments are obviously finite. The majority of current ontology matchers produces non-complex alignment and thus this restriction is only of little impact for the current alignment evaluations. Furthermore, the restriction to consistent aligned ontologies prevents the problems arising from the unsatisfiability as described by Zhang et al. [8]. Given the definition of restricted semantic precision and recall the unsatisfiability of an alignment leads to totally meaningless results for precision and recall and the only reasonable way which does not imply too much additional complexity to the measures is to prohibit the evaluation for inconsistent aligned ontologies. Furthermore, checking for consistency is relatively straight forward using current reasoning tools.

It should be noted that even this restricted version does not comply with the boundedness property demanded by Euzenat which is defined as

$$\mathrm{P}_r(A,R) \geq \mathrm{P}_c(A,R) \wedge \mathrm{R}_r(A,R) \geq \mathrm{R}_c(A,R)$$

where P_c resp. R_c are the classical precision and recall measures. We do not consider this as a real issue. Ehrig and Euzenat [9] introduced this property because the new measures should also consider 'near misses' in a mapping in addition to the exact matches.

Basing the notion of semantic precision and recall on the deductive closure of both, the reference mapping and the generated mapping can naturally lead to an increase as well as a decrease of precision and recall, depending on whether the implied correspondences in the reference mapping outnumber the additional hits derived from the generated mapping or not. The semantically founded evaluation measures should provide the possibility to correctly handle the semantics imminently contained in the ontologies and alignments. Therefore, they are able to raise higher values than the classical measures for alignments which are better from a semantic point of view. The other way around, they should not only provide higher values but also reduce the delivered values compared to the classical measures for alignments which are bad when considering semantics, i.e. imply many wrong correspondences. This is especially important when using those alignments for reasoning, as more implied wrong correspondences lead to more chances to deduce wrong facts from the given knowledge.

IV. Experiments

We implemented a JAVA tool for comparing mappings based on our measures. The tool is able to process OWL ontologies using the OWL API [10] and mappings specified using the alignment format proposed in [11]. The deductive closure is computed using the Pellet reasoner [12]. The Tool is available at http://bitbucket.org/dfleischhacker/spart/. Now we present the results of some experiments conducted using this implementation of restricted semantic precision and recall.

A. Setting

We applied the measures to two different test sets taken from the OAEI test sets. In a first experiment, we applied the measure to the results for the benchmark data set of OAEI 2008, more specifically the ontologies 301 to 304 as these are the most realistic cases from this set. We chose this data set because it is the most commonly used and gives us with the possibility to compare a maximal number of mapping systems.

In a second experiment, we used a subset of the conference set used at the OAEI 2008 consisting of pairwise mappings between five ontologies for which reference alignments exist. This data set was chosen because it features more expressive ontologies and therefore provides a good basis for observing the impact of logical reasoning on the result.

In our experiments we evaluated the alignments as provided by the developers of the ontology matchers which gave us the possibility to assess the restricted measures using realistic automatically generated alignments. In the following, we first present and discuss the results for the benchmark data set before turning our attention to the conference data set.

B. Benchmark Results

We evaluated the alignments submitted to OAEI 2008 by comparing our notion of semantic precision and recall with the values for the classical measures. The results are shown in Table I. The table does not include the results for the natural semantics which fully consist of „err" entries. This is caused by the problem of mappings between objecttype properties and datatype properties mentioned earlier.

The results require some explanation. First of all both alignments CIDER and CIDER-full are provided by the ontology matcher CIDER with different settings. The value „nan" meas that an empty alignment was delivered. Further, the „err" values given for SPIDER in case 304 indicate an error during the execution of our system caused by inappropriate use of namespaces. Thus, we do not have results for these specific cases. Finally, the values are rounded to two digits after the decimal point[2].

On average there was a slight increase of the value for precision. The recall stayed the same for both semantics only in one case, this is for the empty alignment generated by GeRoMe for case 303. As the table depicts, there is no case for which the recall decreased using the new measures. Instead,

[2]The original values are provided at http://dfleischhacker.de/uni/thesis/results.

Testcase / Matchers	Measure	301 P	301 R	302 P	302 R	303 P	303 R	304 P	304 R
aflood	**Classical**	0.96	0.82	0.80	0.60	0.75	0.79	0.95	0.95
	Pragmatic	0.98	0.88	0.87	0.86	0.74	0.90	0.97	0.97
aroma	**Classical**	0.82	0.67	0.76	0.53	0.68	0.56	0.86	0.91
	Pragmatic	0.74	0.71	0.79	0.76	0.57	0.76	0.87	0.94
ASMOV	**Classical**	0.87	0.77	0.56	0.43	0.69	0.79	0.90	0.92
	Pragmatic	0.85	0.83	0.76	0.80	0.74	0.93	0.90	0.95
CIDER	**Classical**	0.88	0.60	0.84	0.55	0.77	0.75	0.93	0.93
	Pragmatic	0.89	0.66	0.94	0.85	0.80	0.86	0.96	0.95
CIDER-full	**Classical**	0.77	0.72	0.61	0.57	0.37	0.75	0.88	0.95
	Pragmatic	0.63	0.81	0.71	0.78	0.28	0.86	0.79	0.97
DSSim	**Classical**	0.87	0.43	0.84	0.55	0.76	0.77	0.96	0.93
	Pragmatic	0.88	0.55	0.78	0.75	0.61	0.90	0.98	0.97
GeRoMe	**Classical**	0.65	0.53	0.31	0.34	nan	0.00	0.75	0.50
	Pragmatic	0.65	0.61	0.37	0.74	nan	0.00	0.78	0.66
Lily	**Classical**	0.91	0.80	0.80	0.60	0.62	0.67	0.94	0.96
	Pragmatic	0.94	0.88	0.78	0.86	0.52	0.77	0.95	0.97
RiMOM	**Classical**	0.75	0.68	0.67	0.62	0.73	0.83	0.89	0.96
	Pragmatic	0.73	0.77	0.72	0.90	0.65	0.97	0.82	0.97
SAMBO	**Classical**	0.94	0.78	0.93	0.57	0.84	0.79	0.95	0.91
	Pragmatic	0.97	0.85	0.98	0.75	0.87	0.89	0.97	0.94
SAMBOdtf	**Classical**	0.94	0.78	0.90	0.57	0.75	0.79	0.93	0.92
	Pragmatic	0.97	0.85	0.96	0.75	0.64	0.90	0.95	0.95
SPIDER	**Classical**	0.24	0.60	0.19	0.55	0.07	0.75	err	err
	Pragmatic	err	err	err	err	err	err	err	err
TaxoMap	**Classical**	1.00	0.22	0.82	0.19	0.58	0.23	0.86	0.32
	Pragmatic	1.00	0.36	0.92	0.20	0.50	0.43	0.84	0.59

TABLE I
RESULTS OF BENCHMARK TEST SET (CLASSICAL - CLASSICAL PRECISION AND RECALL; PRAGMATIC - RESTRICTED SEMANTIC PRECISION AND RECALL WITH PRAGMATIC SEMANTICS)

the recall increased in 47 cases on average by 0.12, which is quite a significant difference. This means that the standard recall measures heavily underestimates the quality of matching systems by not taking implied correspondences into account.

In summary, the increase is much higher for recall than for precision. In the following, we want to have a look at the possible causes.

First, it is to note that the numerator is the same for both measures and it is obvious that $|\mathrm{Cn}(A) \cap \mathrm{Cn}(R)| \geq |A \cap R|$ holds. Thus, the only value which can vary between precision and recall is the denominator which leads to the speculation that the growth of the closure compared to the original alignment is higher for the evaluated alignment than for the reference alignment. To examine this assumption we generated Table II showing the closure growth ratio which is for a given alignment B the value $|\mathrm{Cn}(B)|/|B|$.

This data confirms the assumption that the reference alignment closures growth less than the closure of the alignments for the most cases. This raises the question for the reason of this difference between the closure growth rates. We could assume that the reference alignments should be more complete and already contain many correspondences which can be deduced from other correspondences found in the reference alignment and that the evaluated alignments deduce many more correspondences because they are less complete due to the heuristic nature of the methods used for creating them. A closer look reveals that the reference mappings are minimal and thus contain no redundancy. This fact underlines the need for a semantically founded measure to consider the only implicitly contained knowledge. It seems

Testcase / Matchers	301	302	303	304
Reference	5.47	8.47	4.4	5.55
aflood	5.75	11.29	5.0	5.58
aroma	6.43	11.58	6.98	5.71
ASMOV	6.04	11.69	4.84	5.74
CIDER	5.9	11.61	4.83	5.5
CIDER-full	7.57	10.02	6.65	6.33
DSSim	6.8	12.32	6.31	5.68
GeRoMe	6.24	15.15	empty	7.0
Lily	5.77	12.43	6.0	5.53
RiMOM	6.24	11.63	5.69	6.09
SAMBO	5.8	10.52	4.78	5.58
SAMBOdtf	5.8	10.37	5.8	5.63
TaxoMap	9.0	7.82	9.47	10.5

TABLE II
CLOSURE GROWTH RATIO FOR BENCHMARK TEST CASES USING PRAGMATIC SEMANTICS

Threshold / Matcher	Semantics	0.2 P	0.2 R	0.5 P	0.5 R	0.7 P	0.7 R
ASMOV	*no semantics*	0.42	0.42	0.7	0.18	0.81	0.09
	natural	0.39	0.69	0.81	0.26	1.0	0.15
	pragmatic	0.49	0.74	0.85	0.23	1.0	0.13
DSSim	*no semantics*	0.49	0.52	0.49	0.52	0.49	0.52
	natural	0.15	0.83	0.15	0.83	0.15	0.83
	pragmatic	0.23	0.88	0.23	0.88	0.23	0.88
Lily	*no semantics*	0.5	0.36	0.54	0.21	0.66	0.07
	natural	0.45	0.46	0.65	0.24	0.74	0.09
	pragmatic	0.48	0.51	0.66	0.22	0.65	0.07

TABLE III
AGGREGATED PRECISION (P) AND RECALL (R) RESULTS OF CONFERENCE TEST SET COMPARING CLASSICAL PRECISION AND RECALL (NO SEMANTICS), NATURAL AND PRAGMATIC PRECISION AND RECALL

that most automatically generated alignments contain more correspondences which imply other correspondences than the manually generated reference alignments. So far, we were not able to really explain why this is the case. In future work, we will take a closer look at automatically generated mappings to get a better insight in this phenomenon.

C. Conference Results

Aggregated results for the conference set are presented in Table III. The aggregation is done using the average of all values for a specific measure which are neither an error entry nor have the value „nan". To preserve the comparability if a specific matcher produced an error entry for a specific test case for one semantics, we excluded this matcher-test case combination from the aggregation for each semantics and the classical precision and recall measures. Thus, the tables only contain changes caused by the semantics and not by some side-effects of the semantics regarding their ability to evaluate special correspondences.

The thresholds given in the topmost row reference to the minimum confidence value a correspondence from an alignment must have to be used for the evaluation. Correspondences with a confidence value lower than the threshold are ignored. This behavior causes the values of DSSim to be the same

across all thresholds for a specific semantics as DSSim only has the two confidence levels 0 and 1, where 0 expresses that the matcher considers the correspondence to be invalid.

Some of the most interesting values in the tables are the precision values for the ASMOV matcher for a threshold of 0.7. Using the classical measures the average of all precision values was only 0.81 but using the natural or pragmatic semantics it increased to 1.0. This is because ASMOV provides not only equivalence correspondences but also subsumption relations. The reference alignments only contain equivalence relations and thus the in fact correct subsumption relations raised by ASMOV are falsely considered as incorrect. This is a limitation of the conference data set, because so far all reference alignments only contain equivalences. Hence, at the OAEI only equivalence correspondences are relevant for the precision and recall values which is realized by filtering out all non-equivalence relations. Nevertheless, this aspect shows again the superiority of semantic measures over the classical ones for incomplete reference alignments.

The monotone decrease of the recall and the monotone increase of the precision for higher threshold values are obviously due to the threshold and thus not of interest in regard to this work. The only exception from this property is the precision value of the matcher Lily which decreases for the restricted measure with pragmatic semantics from the threshold 0.5 to the threshold 0.7. This is due to the fact that the alignment generated by Lily for one test case only contains correspondences having a confidence value of less than 0.7. Thus, the threshold 0.7 leads to an empty alignment and therefore the value „nan" for this specific case which has been ignored during the aggregation.

One thing which is possible to assess using the conference test set but not using the benchmark test set is the behavior of the natural and the pragmatic semantics in respect to each other. This is due to the fact that there are alignments which do not contain correspondences between datatype and objecttype properties.

We looked at the growths rates of the closure for the evaluated alignments for both semantics excluding the erroneous results as described above and compared the two. It turns out that the increase is slightly higher for the pragmatic semantics (69 as compared to 64). This means that additional information could be derived using the pragmatic semantics.

In general, the growth rates are much higher for the conference test set than those values for the benchmark test set. This is mainly due to the size of the ontologies and the size of the alignments generated by the matchers. But it is also evident that the values for natural and the pragmatic semantics do not differ by much. This may be due to the fact that the alignments for the conference test set only contain equivalence relations and thus it is not guaranteed that this results holds in general.

V. Related Work

There is some work concerning alternative methods for alignment evaluation. To avoid the problem of not differentiating between full misses and near hits Ehrig and Euzenat [9] introduced a framework for generalizing precision and recall. They instantiated this framework with measures based on the proximity of correspondences given the ontology whereat the proximity values are given by different tables. Furthermore, they gave properties a generalization of precision and recall should fulfill. The main weakness of this table-based approach is the fact that they do not consider the semantics of the ontologies and thus are unable to recognize some near hits. Because of this, Euzenat [4] developed semantically founded variants of precision and recall. For this purpose, he defined the notion of α-consequences of an alignment which are all correspondences deducible from the ontologies given a specific alignment. The ideal measure leads to problems when used for unrestricted alignments due to the possibility to have an infinite set of α-consequences. These weaknesses of the ideal semantic measures were the reason for Euzenat to develop another version of semantically founded measures. Called semantic precision and recall these measures are very similar to the ideal ones but instead of using the α-consequences for both reference and evaluated alignment only the α-consequences of one alignment are used and intersected with the correspondences contained in the other alignment. Thus a possibly infinite set is bound to a set which is guaranteed to be finite. Furthermore, the property of boundedness holds for the semantic precision and recall measures. This definition of semantic precision and recall has some problems. Caused by the use of a pure alignment to bound the set of α-consequences their values depend on the syntax of the alignments. David and Euzenat [5] proposed a normalization of the correspondences contained in the alignments as well as some variants of the ideal measures which use the entirety of all evaluated alignments to bound the sets of α-consequences to bypass these problems.

A related approach is the one presented by Zhang et al. [8] that is based on their framework for purely semantic precision and recall of ontology mappings. This framework provides the possibility to apply different weights to the evaluation of correspondences between concepts, properties and individuals.

VI. Conclusion

We presented a simplified version of semantic precision and recall that restricts the set of implied correspondences to mappings between atomic entities. This restriction makes the measure independent of syntactic variations of the mappings to be compared thus avoiding the problems of previous approaches. We defined the measure and evaluated our implementation of it using results from the Ontology Alignment Evaluation Initiative. Our results show that taking the semantics of the model into account can make a difference in judging the quality of matching systems not only in theory but also in practice. So far, this effect is rather limited, which is mainly due to the fact that most generated alignments as well as reference alignments only consist of equivalence statements. As computing the closure of a mapping mostly adds subsumption relationships to the model, additional overlap between reference and created alignment can only occur in the derived mappings. It is clear, however, that future work will also

strongly focus on generating mappings other than equivalence mappings. Further, there is an ongoing effort to extend existing reference alignments with subsumption correspondences. In such an extended setting, the effect of the semantic measures will be even higher and our system will show its real potential for improving ontology mapping evaluation.

We plan to use the implementation of the measure in future OAEI campaigns. In fact, the 2009 campaign contains a special track on finding subsumption mappings between ontologies. We believe that our system can provide adequate support for evaluating the results of this track. In the longer run, the implementation of precision and recall will be integrated into the SEALS platform, a middleware for supporting the automatic evaluation of semantic technologies that is currently being developed in the EU 7th framework project SEALS.

ACKNOWLEDGEMENTS

This work was partially sponsored by the European Commission in the 7th Framework under contract FP7-238975 (SEALS).

REFERENCES

[1] C. Caracciolo, J. Euzenat, L. Hollink, R. Ichise, A. Isaac, V. Malaisé, C. Meilicke, J. Pane, P. Shvaiko, H. Stuckenschmidt, O. Šváb Zamazal, and V. Svátek, "Results of the Ontology Alignment Evaluation Initiative 2008," in *Proceedings of the ISWC 2008 Workshop on Ontology Matching*, 2008, pp. 73–120.

[2] O. Svab, V. Svatek, and H. Stuckenschmidt, "A Study in Empirical and 'Casuistic' Analysis of Ontology Mapping Results," in *Proceedings of ESWC 2007*, 2007, pp. 655–669.

[3] C. Meilicke and H. Stuckenschmidt, "Incoherence as a basis for measuring the quality of ontology mappings," in *Proceedings of the ISWC 2008 Workshop on Ontology Matching*, 2008, pp. 1–12.

[4] J. Euzenat, "Semantic precision and recall for ontology alignment evaluation," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007, pp. 348–353.

[5] J. David and J. Euzenat, "On fixing semantic alignment evaluation measures," in *Proceedings of the ISWC 2008 Workshop on Ontology Matching*, 2008, pp. 25–36.

[6] C. Meilicke and H. Stuckenschmidt, "An Efficient Method for Computing a Local Optimal Alignment Diagnosis," University of Mannheim, Tech. Rep., 2009.

[7] I. Horrocks, P. Patel-Schneider, and F. van Harmelen, "From SHIQ and RDF to OWL: the making of a Web Ontology Language," *Journal of Web Semantics*, vol. 1, no. 1, pp. 7–26, 2003.

[8] T. Zhang, D. Xu, and J. Chen, "Application-oriented purely semantic precision and recall for ontology mapping evaluation," *Knowledge-Based Systems*, vol. 21, pp. 794–799, 2008.

[9] M. Ehrig and J. Euzenat, "Relaxed precision and recall for ontology matching," in *Proceedings of the K-Cap 2005 Workshop on Integrating Ontology*, 2005, pp. 25–32.

[10] M. Horridge, S. Bechhofer, and O. Noppens, "Igniting the OWL 1.1 Touch Paper: The OWL API," in *Proceedings of the 3rd OWL Experienced and Directions Workshop - OWLED 2007*, Innsbruck, Austria, June 2007.

[11] J. Euzenat, "An API for ontology alignment," in *Proceedings of the 3rd conference on international semantic web conference (ISWC)*, ser. Lecture Notes in Computer Science, D. P. Frank van Harmelen, Sheila McIlraith, Ed., vol. 3298, Hiroshima (JP), 2004, pp. 698–712.

[12] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: A practical OWL-DL reasoner," *Journal of Web Semantics*, vol. 5, no. 2, pp. 51–53, 2007.