Relaxed Precision and Recall for Ontology Matching

Marc Ehrig, Jérôme Euzenat

Relaxed Precision and Recall for Ontology Matching

Marc Ehrig
Institute AIFB
University of Karlsruhe
Karlsruhe, Germany
ehrig@aifb.uni-karlsruhe.de

Jérôme Euzenat
INRIA Rhône-Alpes
655 avenue de l'Europe.
Monbonnot, France
Jerome.Euzenat@inrialpes.fr

ABSTRACT

In order to evaluate the performance of ontology matching algorithms it is necessary to confront them with test ontologies and to compare the results. The most prominent criteria are precision and recall originating from information retrieval. However, it can happen that an alignment be very close to the expected result and another quite remote from it, and they both share the same precision and recall. This is due to the inability of precision and recall to measure the closeness of the results. To overcome this problem, we present a framework for generalizing precision and recall. This framework is instantiated by three different measures and we show in a motivating example that the proposed measures are prone to solve the problem of rigidity of classical precision and recall.

Categories and Subject Descriptors

D.2.12 [**Software**]: Interoperability; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods; D.2.8 [**Software Engineering**]: Metrics

General Terms

Measurement, Performance, Experimentation

Keywords

Ontology alignment, evaluation measures, precision, recall

1. INTRODUCTION

Ontology matching is an important problem for which many algorithms (e.g., PROMPT[11], GLUE[3], Ontrapro[1], OLA[7], FOAM[4]) have been provided. In order to evaluate the performance of these algorithms it is necessary to confront them with test ontologies and to compare the results. The most prominent criteria are precision and recall originating from information retrieval and adapted to the ontology matching task. Precision and recall are based on the comparison of the resulting alignment with another standard alignment, effectively comparing which correspondences are found and which are not. These criteria are well understood and widely accepted.

However, as we have experienced in last year's Ontology Alignment Contest [13], they have the drawback to be of the all-or-nothing kind. An alignment may be very close to the expected result and another quite remote from it and both return the same precision and recall. The reason for this is that the criteria only compare two sets of correspondences without considering if these are close or remote to each other: if they are not the same exact correspondences, they score zero. They both score identically low, despite their different quality. It may be helpful for users to know whether the found alignments are close to the expected one and easily repairable or not. It is thus necessary to measure the proximity between alignments instead of their strict equality.

In this paper we investigate some measures that generalize precision and recall in order to overcome the problems presented above. We first provide the basic definitions of alignments, precision and recall as well as a motivating example (§2). We then present a framework for generalizing precision and recall (§3). This framework is instantiated by four different measures (including classical precision and recall) (§4) and we show on the motivating example that the proposed measures do not exhibit the rigidity of classical precision and recall (§5).

2. FOUNDATIONS

2.1 Alignment

DEFINITION 1 (ALIGNMENT, CORRESPONDENCE). *Given two ontologies O and O', an alignment between O and O' is a set of correspondences (i.e., 4-uples): $\langle e, e', r, n\rangle$ with $e \in O$ and $e' \in O'$ being the two matched entities, r being a relationship holding between e and e', and n expressing the level of confidence $[0..1]$ in this correspondence.*

A matching algorithm returns an alignment A which is compared with a reference alignment R.

K-CAP'05, Workshop Integrating Ontologies, October 2, 2005, Banff, Alberta, Canada.

Let us illustrate this through a simple example. Figure 1 presents two ontologies together with two alignments A_1 and R. In this example, for the sake of simplification, the relation is always '=' and the confidence is always 1.0.

The alignment A_1 is defined as follows:

```
<o1:Vehicle,o2:Thing,=,1.0>
<o1:Car,o2:Porsche,=,1.0>
<o1:hasSpeed,o2:hasProperty,=,1.0>
<o1:MotorKA1,o2:MarcsPorsche,=,1.0>
<o1:250kmh,o2:fast,=,1.0>
```

We present another reasonable alignment A_2:

```
<o1:Car,o2:Thing,=,1.0>
<o1:hasSpeed,o2:hasProperty,=,1.0>
<o1:MotorKA1,o2:MarcsPorsche,=,1.0>
<o1:250kmh,o2:fast,=,1.0>
```

and an obviously wrong alignment A_3:

```
<o1:Object,o2:Thing,=,1.0>
<o1:Owner,o2:Volkswagen,=,1.0>
<o1:Boat,o2:Porsche,=,1.0>
<o1:hasOwner,o2:hasMotor,=,1.0>
<o1:Marc,o2:fast,=,1.0>
```

Further, we have the following reference alignment (R):

```
<o1:Object,o2:Thing,=,1.0>
<o1:Car,o2:Automobile,=,1.0>
<o1:Speed,o2:Characteristic,=,1.0>
<o1:250kmh,o2:fast,=,1.0>
<o1:PorscheKA123,o2:MarcsPorsche,=,1.0>
```

2.2 Precision and Recall

The usual approach for evaluating the returned alignments is to consider them as sets of correspondences and check for the overlap of the two sets. This is naturally obtained by applying the classical measure of precision and recall [14], which are the ratio of the number of true positive ($|R \cap A|$) and retrieved correspondences ($|A|$) or those to be retrieved ($|R|$), respectively.

DEFINITION 2 (PRECISION, RECALL). *Given a reference alignment R, the precision of some alignment A is given by*

$$P(A,R) = \frac{|R \cap A|}{|A|}$$

and recall is given by

$$R(A,R) = \frac{|R \cap A|}{|R|}.$$

2.3 Problems with Current Measures

However, even if the above measurements are easily understandable and widespread, they are often criticized for two reasons: Neither do they discriminate between a totally wrong and an almost correct alignment, nor do they measure user effort to adapt the alignment.

Indeed, it often makes sense to not only have a decision whether a particular correspondence has been found or not, but measure the proximity of the found alignments. This implies that also "near misses" are taken into consideration instead of only the exact matches.

As a matter of example, it will be clear to anybody that among the alignments presented above, A_3 is not a very good alignment and A_1 and A_2 are better alignments. However, they score almost exactly the same in terms of precision (.2) and recall (.2).

Moreover, the alignments will have to go through user scrutiny and correction before being used. It is worth measuring the effort required by the user for correcting the provided alignment instead of only if some correction is needing. This also calls for a relaxation of precision and recall.

3. GENERALIZING PRECISION AND RECALL

Because precision and recall are well-known and easily explained measures, it is good to adhere to them and extend them. It also brings the benefit that measures derived from precision and recall, such as f-measure, can still be computed. For these reasons, we propose to generalize these measures.

If we want to generalize precision and recall, we should be able to measure the proximity of correspondence sets rather than their strict overlap. Instead of the taking the cardinal of the intersection of the two sets ($|R \cap A|$), we measure their proximity (ω).

DEFINITION 3 (GENERALIZED PRECISION AND RECALL). *Given a reference alignment R and an overlap function ω between alignments, the precision of an alignment A is given by*

$$P_\omega(A,R) = \frac{\omega(A,R)}{|A|}$$

and recall is given by

$$R_\omega(A,R) = \frac{\omega(A,R)}{|R|}.$$

3.1 Basic properties

In order, for these new measures to be true generalizations, we would like ω to share some properties with $|R \cap A|$. In particular, the measure should be positive:

$$\forall A,B, \omega(A,B) \geq 0 \qquad \text{(positiveness)}$$

and not exceeding the minimal size of both sets:

$$\forall A,B, \omega(A,B) \leq min(|A|,|B|) \qquad \text{(maximality)}$$

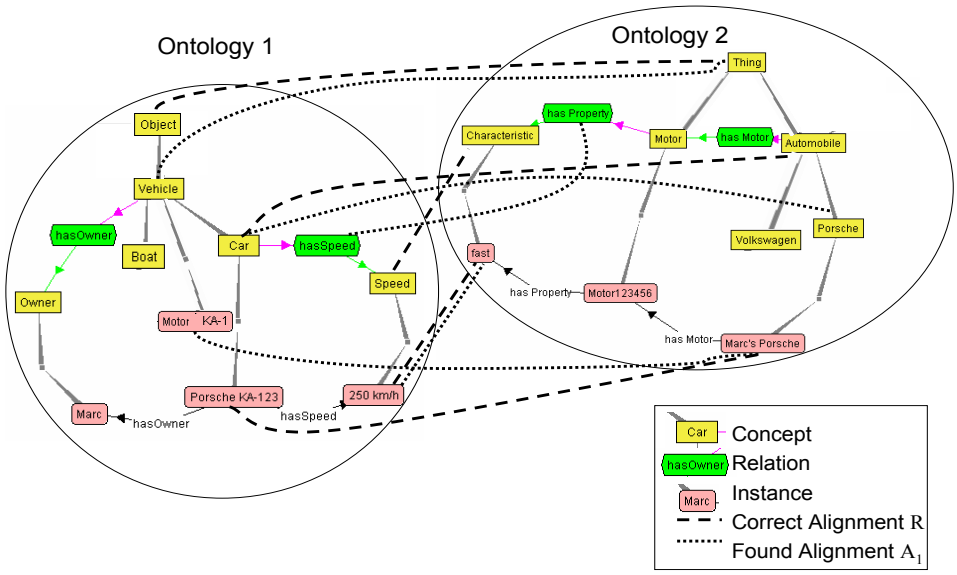

Figure 1: Two Aligned Ontologies

If we want to preserve precision and recall results, ω should only add more flexibility to the usual precision and recall. So their values cannot be worse than the initial evaluation:

$$\forall A, B, \omega(A, B) \geq |A \cap B| \qquad \text{(boundedness)}$$

Hence, the main constraint faced by the proximity is the following:

$$|A \cap R| \leq \omega(A, R) \leq min(|A|, |R|)$$

This is indeed a true generalization because, $|A \cap R|$ satisfies all these properties. One more property satisfied by precision and recall that we will not enforce here is symmetry. This guarantees that the precision and recall measures are true normalized similarities.

$$\forall A, B, \omega(A, B) = \omega(B, A) \qquad \text{(symmetry)}$$

We will not require symmetry, especially since A and R are not in symmetrical positions.

3.2 Designing Overlap Proximity

There are many different ways to design such a proximity given two sets. We retain here the most obvious one which consists of finding correspondences matching each other and computing the sum of their proximity. This can be defined as an overlap proximity:

DEFINITION 4 (OVERLAP PROXIMITY). *A measure that would generalize precision and recall is:*

$$\omega(A, R) = \sum_{\langle a,r\rangle \in M(A,R)} \sigma(a, r)$$

in which $M(A, R)$ is a matching between the correspondences of A and R and $\sigma(a, r)$ a proximity function between two correspondences.

Again, the standard overlap $|A \cap R|$ used in precision and recall is such an overlap proximity.

There are two tasks to fulfill when designing such an overlap proximity function:

- the first one consists of finding the correspondences to be compared M.
- the second one is to define a proximity measure on correspondences σ;

We consider these two issues below.

3.3 Matching Correspondences

A matching between alignments is a set of correspondence pairs, i.e., $M(A, R) \subseteq A \times R$. However, if we want to keep the analogy with precision and recall, it will be necessary to restrict ourselves to the matchings in which an entity from

the ontology does not appear twice. This is compatible with precision and recall for two reasons: (i) in these measures, any correspondence is identified only with itself, and (ii) appearing more than once in the matching would not guarantee an overlap proximity below $min(|A|, |R|)$.

There are $\frac{|A|!}{(|A|-|R|)!}$ candidate matches (if $|A| \geq |R|$). The natural choice is to select the best match because this guarantees that the function generalizes precision and recall.

DEFINITION 5 (BEST MATCH). *The best match $M(A, R)$ between two sets of correspondences A and R, is the subset of $A \times R$ which maximizes the overall proximity and in which each element of A (resp. R) belongs to only one pair:*

$$M(A, R) \in Max_{\omega(A,R)}\{M \subseteq A \times R\}$$

As defined here, this best match may not be unique. This is not a problem, because we only want to find the highest value for ω and any of the best matches will yield the same value.

Of course, the definitions M and ω are dependent of each other, but this does not prevent us from computing them. They are usually computed together but it is better to present them separately.

3.4 Correspondence Proximity

In order to compute $\omega(A, R)$, we need to measure the proximity between two matched correspondences (i.e., $\langle a, r\rangle \in M(A, R)$) on the basis of how close the result is from the ideal one. Each element in the tuple $a = \langle e_a, e'_a, r_a, n_a\rangle$ will be compared with its counterpart in $r = \langle e_r, e'_r, r_r, n_r\rangle$. For any two correspondences (the found a and the reference r) we compute three similarities σ_{pair}, σ_{rel}, and σ_{conf}. If elements are identical, proximity has to be one (maximality). If they differ, proximity is lower, always according to the chosen strategy. In contrast to the standard definition of similarity, the mentioned proximity measures do not necessarily have to be symmetric. We will only consider normalized proximities, i.e., measures whose values are within the unit interval $[0..1]$, because this guarantees that

$$\omega(A, R) \leq min(|A|, |R|)$$

The component proximity measure is defined in the following way:

$\sigma_{pair}(\langle e_a, e_r\rangle, \langle e'_a, e'_r\rangle)$: How is one entity pair similar to another entity pair? In ontologies we can in principal follow any relation which exists (e.g., subsumption, instantiation), or which can be derived in a meaningful way. The most important parameters are the relations to follow and their effect on the proximity.

$\sigma_{rel}(r_a, r_r)$: Often the alignment relations are more complex, e.g., represent subsumption, instantiation, or compositions. Again, one has to assess the similarity between these relations. The two relations of the alignment cell can be compared based on their distance in a conceptual neighborhood structure [6, 8].

$\sigma_{conf}(n_a, n_r)$: Finally, one has to decide, what to do with different levels of confidence. The similarity could simply be the difference. Unfortunately, none of the current alignment approaches have an explicit meaning attached to confidence values, which makes it rather difficult in defining an adequate proximity.

Once these proximities are established, they have to be aggregated. The constraints on the aggregation function ($Aggr$) are:

- normalization preservation (if $\forall i, 0 \leq c_i \leq 1$ then $0 \leq Aggr_i c_i \leq 1$);
- maximality (if $\forall i, c_i = 1$ then $Aggr_i c_i = 1$);
- local monotonicity (if $\forall i \neq j, c_i = c'_i = c''_j$ and $c_j \leq c'_j \leq c''_j$ then $Aggr_i c_i \leq Aggr_i c'_i \leq Aggr_i c''_i$).

Here, we consider aggregating them through multiplication without further justification. Other aggregations (e.g., weighted sum) are also possible.

DEFINITION 6 (CORRESPONDENCE PROXIMITY). *Given two correspondences $\langle e_a, e'_a, r_a, n_a\rangle$ and $\langle e_r, e'_r, r_r, n_r\rangle$, their proximity is:*

$$\sigma(\langle e_a, e'_a, r_a, n_a\rangle, \langle e_r, e'_r, r_r, n_r\rangle) =$$

$$\sigma_{pair}(\langle e_a, e_r\rangle, \langle e'_a, e'_r\rangle) \times \sigma_{rel}(r_a, r_r) \times \sigma_{conf}(n_a, n_r)$$

We have provided constraints and definitions for M, ω, and σ. We now turn to concrete measures.

4. CONCRETE MEASURES

We consider four cases of relaxed precision and recall measures based on the above definitions. We first give the definition of usual precision and recall within this framework.

4.1 Standard Precision and Recall

For standard precision and recall, the value of ω is $|A \cap R|$. This is indeed an instance of this framework, if the proximity used is based on the strict equality of the components of correspondences.

DEFINITION 7 (EQUALITY PROXIMITY). *The equality*

proximity is charaterized by:

$$\sigma_{pair}(\langle e_a, e'_a\rangle, \langle e_r, e'_r\rangle) = \begin{cases} 1 & \text{if } \langle e_a, e'_a\rangle = \langle e_r, e'_r\rangle \\ 0 & \text{otherwise} \end{cases}$$

$$\sigma_{rel}(r_a, r_r) = \begin{cases} 1 & \text{if } r_a = r_r \\ 0 & \text{otherwise} \end{cases}$$

$$\sigma_{conf}(n_a, n_r) = \begin{cases} 1 & \text{if } n_a = n_r \\ 0 & \text{otherwise} \end{cases}$$

4.2 Symmetric Proximity

The easiest way to relax precision and recall is to have some distance δ on the elements in ontologies and to weight the proximity with the help of this distance: the higher the distance between two entities in the matched correspondences, the lower their proximity. This can be defined as:

$$\left.\begin{array}{rl} & \delta(e_a, e_r) \leq \delta(e_b, e_r) \\ \text{and} & \delta(e'_a, e'_r) \leq \delta(e'_b, e'_r) \end{array}\right\}$$

$$\Longrightarrow \sigma(\langle e_a, e'_a\rangle, \langle e_r, e'_r\rangle) \geq \sigma(\langle e_b, e'_b\rangle, \langle e_r, e'_r\rangle)$$

As a simple example of such a symmetric similarity, we use a distance in which a class is at distance 0 of itself, at distance 0.5 of its direct sub- and superclasses, and at a distance 1 of any other class. This could be further refined by having a similarity inversely proportional to the distance in the subsumption tree. Likewise, this similarity may also be applied to properties and instances (through part-of relationships in the latter case). The similarity between pairs is the complement of these similarities The result is displayed in Table 1. We always mention the assumed alignment and the actual correct alignment.

found e,e'	closest correct e,e'	similarity σ_{pair}	comment
e,e'	e,e'	1	correct correspondence
c,c'	c,$sup(c')$	0.5	returns more specialized instances
c,c'	$sup(c)$,c'	0.5	returns more general instances
c,c'	c,$sub(c')$	0.5	returns more general instances
c,c'	$sub(c)$,c'	0.5	returns more specialized instances
r,r'	r,$sup(r')$	0.5	returns more spec. relation instances
r,r'	$sup(r)$,r'	0.5	returns more gen. relation instances
r,r'	r,$sub(r')$	0.5	returns more gen. relation instances
r,r'	$sub(r)$,r'	0.5	returns more spec. relation instances
i,i'	i,$super(i')$	0.5	returns a more restricted instance
i,i'	$super(i)$,i'	0.5	returns a too broad instance
i,i'	i,$sub(i')$	0.5	returns a too broad instance
i,i'	$sub(i)$,i'	0.5	returns a more restricted instance

Table 1: Similarities based on Entity Pairs

Table 2 consider the proximity between relations. It only presents the similarity between equality (=) and other relations.

For the confidence distance we simply take the complement of the difference. The final precision is calculated according to the formula presented in the previous section:

found relation	correct relation	similarity σ_{rel}	comment
$e = e'$	$e = e'$	1	correct relation
$c = c'$	$c \subset c'$	0.5	returns more instances than correct
$c = c'$	$c \supset c'$	0.5	returns less instances than possible, but these are correct
$r = r'$	$r \subset r'$	0.5	
$r = r'$	$r \supset r'$	0.5	
$i = i'$	i partOf i'	0.5	
$i = i'$	i consistsOf i'	0.5	

Table 2: Similarities based on Relations

DEFINITION 8 (SYMMETRIC PROXIMITY). *The symmetric proximity is characterized by:*

$$\sigma_{pair}(\langle e_a, e'_a\rangle, \langle e_r, e'_r\rangle) \text{ as defined in Table 1}$$
$$\sigma_{rel}(r_a, r_r) \text{ as defined in Table 2}$$
$$\sigma_{conf}(n_a, n_r) = 1 - |n_a - n_r|.$$

4.3 Measuring Correction Effort

If users have to check and correct alignments, the quality of alignment algorithms can be measured through the effort required for transforming the obtained alignment into the (correct) reference one [2].

This measure can be implemented as an edit distance [10]: an edit distance defines a number of operations by which an object can be corrected (here the the operations on correspondences authorized) and assigns a cost to each of these operations (here the effort required to identify and repair some mistake). The cost of a sequence of operations is the sum of their cost and the distance between two objects is the cost of the less costly sequence of operations that transform one object into the other one. The result can always be normalized in function of the size of the largest object. Such a distance can be turned into a proximity by taking its complement with regard to 1.

Table 3 provides such plausible weights. Usually classes are organized in a taxonomy in which they have less direct super- than subclasses. It is thus easier to correct a class to (one of) its superclass than to one of its subclasses. As a consequence, the proximity is dissymmetric. Such a measure should also add some effort when classes are not directly related, but this has not been considered here.

The edit distance between relations is relatively easy to design since, generally, changing from one relation to another can be done with just one click. Thus, the relational similarity equals 1 if the relations are the same and 0.5 otherwise.

In this correction effort measure, the confidence factor does not play an important role: ordering the correspondences can only help the user to know that after some point she will have to discard many correspondences. We thus decided to not take confidence into account and thus, their proximity will always be 1.

found e,e'	closest correct e,e'	effort	similarity σ_{pair}	comment
e,e'	e,e'	0	1	correct alignment
c,c'	$c,sup(c')$	0.4	0.6	returns more spec. instances
c,c'	$sup(c),c'$	0.4	0.6	returns more gen. instances
c,c'	$c,sub(c')$	0.6	0.4	returns more gen. instances
c,c'	$sub(c),c'$	0.6	0.4	returns more spec. instances
r,r'	$r,sup(r')$	0.4	0.6	
r,r'	$sup(r),r'$	0.4	0.6	
r,r'	$r,sub(r')$	0.6	0.4	
r,r'	$sub(r),r'$	0.6	0.4	
i,i'	$i,super(i')$	0.4	0.6	returns a more restricted inst.
i,i'	$super(i),i'$	0.4	0.6	returns a too broad inst.
i,i'	$i,sub(i')$	0.6	0.4	returns a too broad inst.
i,i'	$sub(i),i'$	0.6	0.4	returns a more restricted inst.

Table 3: Effort-based proximity between Entity Pairs

DEFINITION 9 (EFFORT-BASED PROXIMITY). *The effort-based proximity is charaterized by:*

$\sigma_{pair}(\langle e_a, e'_a\rangle, \langle e_r, e'_r\rangle)$ *as defined in Table 3*

$$\sigma_{rel}(r_a, r_r) = \begin{cases} 1 & \text{if } r_a = r_r \\ 0.5 & \text{otherwise} \end{cases}$$

$$\sigma_{conf}(n_a, n_r) = \begin{cases} 1 & \text{if } n_a \neq 0 \text{ and } n_r \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

To be accurate, such an effort proximity would have been better aggregated with an additive and normalized aggregation function rather than multiplication.

4.4 Precision- and Recall-oriented Measures

One can also decide to use two different similarities depending on their application for evaluating either precision or recall. We here provide two such measures and justify the given weights. Precision is normally a measure of accuracy i.e., the returned results need to be correct. Every wrong result will therefore entail a penalty. We assume the user poses a query to the system as follows: "return me all instances of e". The system then returns any instance corresponding to the alignment i.e. e'. Vice versa, for the relaxed recall we want to avoid missing any correct result. This affects the similarity relations and weights.

4.4.1 Relaxed Precision

In Table 4 and 5 we present the precision similarity for pairs and relations. The comments in each line explain the decision for the weights.

For the distance within the confidence we again use the complement of the difference.

DEFINITION 10 (PRECISION-ORIENTED PROXIMITY). *The precision-recall oriented proximity is characterized by:*

$\sigma_{pair}(\langle e_a, e'_a\rangle, \langle e_r, e'_r\rangle)$ *as defined in Table 4*

$\sigma_{rel}(r_a, r_r)$ *as defined in Table 5*

$\sigma_{conf}(n_a, n_r) = 1 - |n_a - n_r|$.

found e,e'	closest correct e,e'	similarity σ_{pair}	comment
e,e'	e,e'	1	correct correspondence
c,c'	$c,sup(c')$	1	returns more specialized instances, these are correct
c,c'	$sup(c),c'$	0.5	returns more general instances, includes some correct results
c,c'	$c,sub(c')$	0.5	returns more general instances, includes some correct results
c,c'	$sub(c),c'$	1	returns more specialized instances, these are correct
r,r'	$r,sup(r')$	1	
r,r'	$sup(r),r'$	0.5	
r,r'	$r,sub(r')$	0.5	
r,r'	$sub(r),r'$	1	
i,i'	$i,super(i')$	0.5	returns a more restricted instance
i,i'	$super(i),i'$	0	returns a too broad instance
i,i'	$i,sub(i')$	0	returns a too broad instance
i,i'	$sub(i),i'$	0.5	returns a more restricted instance

Table 4: Similarities for Relaxed Precision based on Entity Pairs

found relation	correct relation	similarity σ_{rel}	comment
$e = e'$	$e = e'$	1	correct relation
$c = c'$	$c \subset c'$	0.5	returns more instances than correct
$c = c'$	$c \supset c'$	1	returns less instances than possible, but these are correct
$r = r'$	$r \subset r'$	0.5	
$r = r'$	$r \supset r'$	1	
$i = i'$	i partOf i'	0.5	
$i = i'$	i consistsOf i'	1	

Table 5: Similarities for Relaxed Precision based on Relations

4.4.2 Relaxed Recall

In Table 6 and 7 we present the recall similarity for pairs and relations. Basically many distances are just mirrored compared to the precision case.

found e,e'	closest correct e,e'	similarity σ_{pair}	comment
e,e'	e,e'	1	correct correspondence
c,c'	$c,sup(c')$	0.5	returns more specialized instances, misses some
c,c'	$sup(c),c'$	1	returns more general instances, includes the correct results
c,c'	$c,sub(c')$	1	returns more general instances, includes the correct results
c,c'	$sub(c),c'$	0.5	returns more specialized instances, misses some
r,r'	$r,sup(r')$	0.5	
r,r'	$sup(r),r'$	1	
r,r'	$r,sub(r')$	1	
r,r'	$sub(r),r'$	0.5	
i,i'	$i,super(i')$	0	returns a more restricted instance, misses correct
i,i'	$super(i),i'$	0.5	returns a broader instance
i,i'	$i,sub(i')$	0.5	returns a broader instance
i,i'	$sub(i),i'$	0	returns a more restricted instance, misses correct

Table 6: Similarities for Relaxed Recall based on Entity Pairs

found relation	correct relation	similarity σ_{rel}	comment
$e = e'$	$e = e'$	0	correct relation
$c = c'$	$c \subset c'$	0	returns more instances than correct
$c = c'$	$c \supset c'$	0.5	returns less instances than possible, misses some
$r = r'$	$r \subset r'$	0	
$r = r'$	$r \supset r'$	0.5	
$i = i'$	i partOf i'	0	
$i = i'$	i consistsOf i'	0.5	

Table 7: Similarities for Relaxed Recall based on Relations

The final recall is computed as usual:

DEFINITION 11 (RECALL-ORIENTED PROXIMITY). *The recall-oriented proximity is characterized by:*

$$\sigma_{pair}(\langle e_a, e'_a\rangle, \langle e_r, e'_r\rangle) \text{ as defined in Table 6}$$
$$\sigma_{rel}(r_a, r_r) \text{ as defined in Table 7}$$
$$\sigma_{conf}(n_a, n_r) = 1 - |n_a - n_r|.$$

5. EXAMPLE

In the introduction of this paper we have presented a pair of ontologies, the reference alignment, and three different identified alignments. We will now apply the different proposed precision and recall measures to these example alignments. Please note that they mainly illustrate entity pair similarities, as relations and confidences are always identical. Table 8 provides the results. For the oriented measure we assume that the query is given in ontology 1 and the answer has to be retrieved in ontology 2. As the oriented measure is dissymmetric, one has to define this direction beforehand.

ω	(R, R)		(R, A_1)		(R, A_2)		(R, A_3)	
	P	R	P	R	P	R	P	R
standard	1.0	1.0	0.2	0.2	0.25	0.2	0.2	0.2
symmetric	1.0	1.0	0.4	0.4	0.375	0.3	0.2	0.2
edit	1.0	1.0	0.44	0.44	0.35	0.28	0.2	0.2
oriented	1.0	1.0	0.5	0.5	0.375	0.4	0.2	0.2

Table 8: Precision recall result on the alignments of Figure 1

The measures which have been introduced address the problems raised in the introduction and fulfill the requirements:

- They keep precision and recall untouched for the best alignment (R);
- They help discriminating between irrelevant alignments (A_3) and not far from target ones (A_1 and A_2);
- Specialized measures are able to emphasize some characteristics of alignments: ease of modification, correctness or completeness. For instance, let's consider the oriented measures. In our example A_1 has two very near misses, which leads to a relatively high precision. In A_2 however the miss is bigger, but by aligning one concept to its superconcept recall rises relatively to precision.

These results are based on only one example. They have to be systematized in order to be extensively validated. Our goal is to implement these measures within the Alignment API and to use them on the forthcoming results of the Ontology Alignment Evaluation 2005[1] in order to have real data on which the relevance of the proposed measures can be more openly debated.

6. RELATED WORK

The naturally relevant work is [2] which has considered precisely the evaluation of schema matching. However, the authors only note the other mentioned problem (having two measures instead of one) and use classical aggregation (overall and F-measure) of precision and recall.

In computational linguistics, and more precisely multilingual text alignment, [9] has considered extending precision and recall. Their goal is the same as ours: increasing the discriminating power of the measures. In this work, the mathematical formulation is not changed but the granularity of compared sets changes: instead of comparing sentences in a text, they compare words in sentences in a text. This helps having some contribution to the measures when most of the words are correctly aligned while the sentences are not strictly aligned.

In the Alignment API [5], there is another evaluation measure which directly computes a distance based on a weighted symmetric difference (weights are the confidences of each correspondence in the alignment). This measure could be used in the generalization proposed here (the distance would then be based on confidence difference and would generally satisfy $P'(A, R) \leq P(A, R)$ and $R'(A, R) \leq R(A, R)$.

The deeper proposal for extending precision and recall comes from hierarchical text categorization in which texts are attached to some category in a taxonomy [12]. Usually, texts are attached to the leaves, but when algorithms attach them to the intermediate categories, it is useful to discriminate between a category which is irrelevant and a category which is an immediate super category of the expected one. For that purpose, they introduce an extension of precision (recall is redefined similarly) such that:

$$P_{CS} = \frac{max(0, |A \cap R| + FpCon + FnCon)}{|A| + FnCon}$$

in which $FpCon$ (resp. $FnCon$) is the contribution to false positive (resp. false negative), i.e., the way incorrectly classified documents could contribute to its incorrect category anyway. The maximization is necessary to prevent the result from being negative (because the contribution is defined with respect to the average such contribution). The contribution is measured in two ways. The first one is a category similarity that is computed on the features of categories (categories and documents are represented by a vector of features and the membership to some category is based on a distance be-

[1] http://oaei.inrialpes.fr/2005/

tween these vectors). The second one is based on the distance between categories in the taxonomy.

This measure does not seem to be a generalization of standard precision and recall as the one presented here. In particular, because the contributions can be negative, this measure can be lower than standard precision and recall. The idea of retracting the contribution from wrongly classified documents is not far from the idea developed here. However, the computation of this contribution with regard to some average and the addition of some contribution to the divisor do not seem justified.

7. DISCUSSION

Evaluation of matching results is often made on the basis of the well-known and well-understood precision and recall measures. However, these measures do not discriminate accurately between methods which do not provide the exact results. In the context where the result of alignments have to be screened by humans, this is an important need.

We have proposed a framework for generalizing precision and recall when comparing ontology alignments. It keeps the advantages of usual precision and recall but helps discriminating between alignments by identifying for near misses instead of completely wrong correspondences.

The framework has been instantiated in three different measures, each one aiming at favoring some particular aspects of alignment utility. We show that these measures indeed avoid the shortcomings of standard evaluation criteria. They should however, be further investigated in order to find better formulations: more discrepancy needs to be considered, more progressive distance (e.g., not direct subclasses) and rationalized design of weights.

This generalization framework is not the only possible one since we have made a number of choices:

- on the form of the alignment similarity (Definition 4);
- on the kind of alignment matching (Definition 5);
- on the form of the correspondence similarity (Definition 6).

More work has to be done in order to assess the potential of other choices in these functions.

The most important work is to consider these proposed measures in real evaluation of alignment systems and to identify good measures for further evaluations. We plan to implement these measures within the Alignment API [5] and process the results of the Ontology Alignment Evaluation 2005.

Acknowledgements

This work has been partially supported by the Knowledge Web European network of excellence (IST-2004-507482). The authors would like to thank Diana Maynard who pointed out the problem addressed here.

8. REFERENCES

[1] B. Ashpole. Ontology translation protocol (ontrapro). In E. Messina and A. Meystel, editors, *Proceedings of Performance Metrics for Intelligent Systems (PerMIS '04)*, Gaithersburg, MD, USA, August 2004.

[2] H.-H. Do, S. Melnik, and E. Rahm. Comparison of schema matching evaluations. In *Proc. GI-Workshop "Web and Databases", Erfurt (DE)*, 2002. http://dol.uni-leipzig.de/pub/2002-28.

[3] A.-H. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy. Learning to map ontologies on the semantic web. *VLDB journal*, 2003.

[4] M. Ehrig and S. Staab. QOM - quick ontology mapping. In *Proc. 3rd ISWC, Hiroshima (JP)*, November 2004.

[5] J. Euzenat. An API for ontology alignment. In *Proc. 3rd international semantic web conference, Hiroshima (JP)*, pages 698–712, 2004.

[6] J. Euzenat, N. Layaïda, and V. Dias. A semantic framework for multimedia document adaptation. In *Proc. 18th International Joint Conference on Artificial Intelligence (IJCAI), Acapulco (MX)*, pages 31–36, 2003.

[7] J. Euzenat and P. Valtchev. Similarity-based ontology alignment in OWL-lite. In *Proc. 15th ECAI*, pages 333–337, Valencia (ES), 2004.

[8] C. Freksa. Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54(1–2):199–227, 1992.

[9] P. Langlais, J. Véronis, and M. Simard. Methods and practical issues in evaluating alignment techniques. In *Proc. 17th international conference on Computational linguistics, Montréal (CA)*, pages 711–717, 1998.

[10] I. V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 1966.

[11] N. Noy and M. Musen. Smart: Automated support for ontology merging and alignment, 1999.

[12] A. Sun and E.-P. Lin. Hierarchical text classification and evaluation. In *Proc. IEEE international conference on data mining*, pages 521–528, 2001.

[13] Y. Sure, O. Corcho, J. Euzenat, and T. Hughes, editors. *Proceedings of the 3rd Evaluation of Ontology-based tools (EON)*, 2004.

[14] C. J. van Rijsbergen. *Information retrieval*. Butterworths, London (UK), 1975. http://www.dcs.gla.ac.uk/Keith/Preface.html.