Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

RECOMMENDATIONS FOR BETTER QUALITY ONTOLOGY MATCHING EVALUATIONS

Aliaksandr Autayeu, Vincenzo Maltese and Pierre Andrews

March 2010

Technical Report # DISI-10-024

Also: in proceedings of the AISB Workshop on Matching and Meaning 2010, 31st March 2010, Leicester, UK

Recommendations for Better Quality Ontology Matching Evaluations[1]

Aliaksandr Autayeu and **Vincenzo Maltese** and **Pierre Andrews** [2]

Abstract. Evaluating and comparing different ontology matching techniques is a complex multifaceted problem. Currently, diverse golden standards and various practices are used for evaluations. In this paper we show that, by following certain rules, the quality of the evaluations can be significantly improved, particularly in regard to the accuracy of precision and recall measures obtained.

1 Introduction

In the recent years, several matching tools have been proposed as solutions to the semantic heterogeneity problem. We focus on the problem of evaluating semantic matching techniques [10] which find rich mappings, that is, mappings which contain *disjointness*, *equivalence*, *less general* and *more general* links.

Most of the tools for semantic matching identify only equivalence, some identify less and more generality, but only a few include explicit disjointness [23]. Reflecting this, an overwhelming majority of available golden standards are targeted at evaluating mappings with equivalence links only. In this paper we explain why evaluating a rich mapping using such a golden standard makes results imprecise and propose some recommendations to make evaluations and comparison between different tools fairer and more accurate.

In addition, we address important issues concerning the coverage of a golden standard and the presence of redundant links, namely links which can be logically inferred from the others in the mapping, and how they can influence the quality measures.

The rest of the paper is organized as follows. In Section 2 we present state of the art approaches. Section 3 shows how to conduct the evaluation according to the characteristics of the golden standard used. Section 4 focuses on the importance of maximizing both the mapping and the golden standard for an optimal evaluation. Finally, Section 5 concludes the paper and outlines future work.

2 Related Work

With matching techniques being the main focus of the ontology matching field, a few initiatives[3] pay attention to evaluation. On the one hand, general [22, 6] and domain-specific [18, 17] evaluation experiments are reported, without discussing the evaluation methodology. On the other hand, considerable attention has been paid to appropriateness and quality of the measures [8, 9, 4].

Attention has also been brought to the mapping itself. In [20] the authors propose to complement the precision and recall with new measures to take into account possible mapping incoherence, thus addressing the issues of internal logical problems of the mapping and the lack of reference mappings. In [24] two evaluation techniques are proposed. The first is practice-oriented and evaluates the behaviour of the mapping in use. The second focuses on the manual evaluation of a mapping sample and the generalization of the results.

Closer to our work, Sabou and Gracia [23] raise the issue of evaluating non-equivalence links, pointing out that several systems also produce subsumption and disjointness links. In particular, they discuss the issue of evaluating a mapping that contains redundant links, that is, links that can be logically derived from the others in the mapping. They compute precision both for the original set and the set from which the redundant links are removed. We extend and improve their conclusions.

3 Computing Precision and Recall

Golden standards are fundamental for computing the well-known precision and recall measures [4]. Typically, hand-made positive (GS^+) and negative (GS^-) golden standards contain links considered correct and incorrect, respectively. Ideally, GS^- complements GS^+, leading to a precise evaluation. Yet, annotating all links in big datasets (with thousands or millions of links) is impossible and therefore the golden standard is often composed of three sets:

$\boldsymbol{GS^+}$ the set of links considered correct;
$\boldsymbol{GS^-}$ the set of links considered incorrect;
$\boldsymbol{Unk}$ the pairs of nodes for which the semantic relation is unknown.

If we denote the result of the matcher (the mapping) with Res, precision and recall can be computed as follows [14]:

$$Precision = \frac{TP}{TP + FP} = \frac{|Res \cap GS^+|}{|Res \cap GS^+| + |Res \cap GS^-|} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} = \frac{|Res \cap GS^+|}{|GS^+|}, \quad (2)$$

where:

$\boldsymbol{TP}$ (True Positives) is the set of links found by the algorithm that hold;
$\boldsymbol{FP}$ (False Positives) is the set of links found by the algorithm that do not hold;
$\boldsymbol{FN}$ (False Negatives) is the set of links that hold, but which were not found by the algorithm.

For the cases in which GS^- is not available, precision can be approximated as follows:

$$Precision = \frac{|Res \cap GS^+|}{|Res|} \quad (3)$$

[1] This is an extended version of the TechRep DISI-09-046, also presented as a poster at the ISWC Ontology Matching Workshop (OM 2009), 25th October 2009, Washington DC, USA

[2] University of Trento, Italy, email: surname@disi.unitn.it

[3] http://oaei.ontologymatching.org/

These sets are illustrated in Figure 1. For example, if for sake of simplicity we use numbers to indicate links, we could have:

$$Res = \{1,2,3,4\} \qquad Unk = \{\}$$
$$GS^- = \{3,4,6,8\} \quad GS^+ = \{1,2,5,7,9,10\}$$
$$Precision = \frac{2}{(2+2)} = 0.5 \qquad Recall = \tfrac{2}{6} = 0.33 \qquad (4)$$

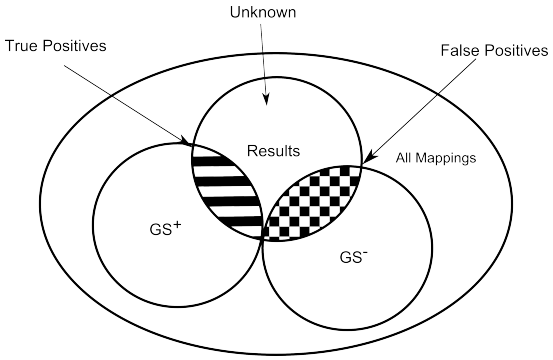


Figure 1. True Positives, False Positives and Golden Standards

The precision gives an indication of the amount of noise that is retrieved by the matching algorithm (how many correct links it returns) while the recall is a measure of the coverage of the algorithm (how many links the algorithm found and missed).

3.1 Coverage of the Golden Standard

Given two ontologies of size n and m, the size of a mapping and the golden standards range in $[0, n \times m]$. To enable precise computation of precision and recall, one should inspect all $n \times m$ combinations of nodes and consider all possible semantic relations that can hold between them. For large ontologies this is practically impossible. The huge effort required for their construction is the main reason why only a few golden standards are available and evaluation campaigns tend to use very small ontologies, risking a loss of statistical significance of the results and biasing towards one algorithm or the other.

When setting up exhaustive GS^+ and GS^- is not possible, the common practice is to inspect only a subset of the $n \times m$ node pairs [2, 14]. Partial coverage leads to an approximated evaluation. In particular, we cannot say anything about the subset $Res \cap Unk$ of the links. However, if GS^+ and GS^- are sampled properly, the precision and recall should still be evaluated in a statistically significant manner. For example, we could have reduced coverage compared to the previous example, as follows:

$$Res = \{1,2,3,4\} \quad GS^+ = \{1,2,7\} \quad GS^- = \{3,6,8\}$$
$$Unk = \{4,5,9,10\} \quad Res \cap Unk = \{4\}$$
$$Precision = \frac{2}{(2+1)} = 0.66 \quad Recall = \frac{2}{3} = 0.66$$

As it turns out (compare the measures with those in (4)), such evaluations may be very different from the real values, therefore:

Recommendation 1. Use large golden standards. Include GS^- for a good approximation of the precision and recall. To be statistically significant, cover in GS^+ and GS^- an adequate portion of all node pairs.

In a sampled GS, reliability of results depends on: (a) the portion of the pairs covered by the golden standard; (b) the ratio between the size of the GS^+ and the GS^- and (c) their quality (see Section 4).

3.2 Comparing Semantic Relations

Current state of the art tools output different kinds of relations. While most of the matching tools, such as Similarity Flooding [21], Cupid [19] and COMA [7] only produce *equivalence*, some tools, such as AROMA [5] also produce *less general* and *more general* relations. At the best of our knowledge only ctxMatch [3], S-Match [12], Min S-Match [11] and Spider [23] also produce explicit *disjointness*.

Currently, different tools are usually compared without distinguishing among the different semantic relations produced and only the presence or absence of a relation between a pair of nodes is evaluated. This means, for instance, that subsumption and equivalence are considered as the same. This approach can be used to compare heterogeneous mappings, but leads to imprecise results.

A particular discourse has to be made for *disjointness* relations. Typically disjointness links are seen as a negative result, that is, a clear indication of two completely unrelated nodes. Thus, the majority of matching tools do not consider them interesting to the users. As a consequence they do not compute them at all, but corresponding node pairs are rather put in the GS^-. Moreover, they are often confused with *overlap*.

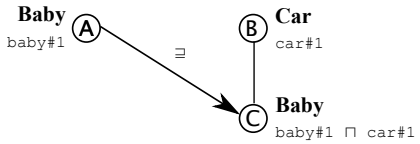


Figure 2. Overlap between nodes A and B. Natural language labels are in bold with a corresponding DL formulas under them.

Consider the example in Figure 2. The link $< A, C, \sqsupseteq >$ is a correct result and as such should be part of the GS^+. In fact, given the semantics of lightweight ontologies [16], the meaning of the node C includes the meaning of the node B above it. What about the relation between A and B then? They are not disjoint as they share C. The relation is rather an *overlap* (namely $A \cap B \neq \emptyset$).

Discriminating the two cases above is fundamental both to conclude the right relations between the nodes and to correctly evaluate precision and recall of disjointness relations when they are explicitly computed by the matching tool. In fact, the main problem is that negative golden standards (when available) typically contain undifferentiated links. For instance, the authors of [14] make no difference between *disjointness* and *overlap* relations. To the best of our knowledge, no evaluations take *disjointness* and *overlap* relations into account when measuring precision and recall. To summarize:

Recommendation 2. When presenting evaluation results, specify whether and how the evaluation takes into account the kinds of the semantic relations.

4 Maximized and Minimized Golden Standards

We use the notion of minimal mapping [11] to judge the quality of a golden standard. The basic idea is that among all possible links between two ontologies there are some redundant ones, which can be logically inferred from the others. The minimal mapping is defined

as the minimal set of (non-redundant) links such that all the other (redundant) links can be logically inferred from the non-redundant ones [11, 13]. We use the **Min(mapping)** function to remove the redundant links from the mapping (producing the *minimized mapping*) and the **Max(mapping)** function to add all the redundant links (producing the *maximized mapping*).

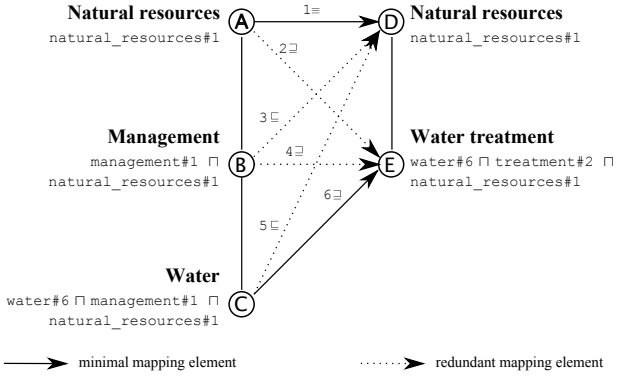


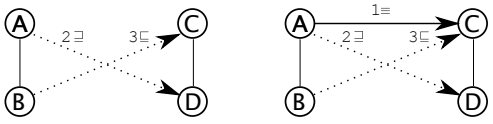
Figure 3. The mapping between two lightweight ontologies. Original natural language labels are in bold.

Consider the example in Figure 3 taken from [13]. It shows the minimal mapping (the solid arrows) and the mapping of maximum size (including the maximum number of redundant links, represented as dashed arrows) computed between two lightweight ontologies.

In the following we provide three observations.

The first observation is that following [11] and staying within lightweight ontologies guarantees that the maximized mapping is always finite and thus corresponding precision and recall can always be computed.

The second observation is that, in contrast with [23], we argue (and show with an example) that comparing the minimized versions of the mapping and the golden standards is not informative. The reason is that minimization can significantly reduce the amount of links in their intersection. In other words, they can share a few non-redundant links still generating a significant amount of redundant links in common. Notice that different non-redundant links can generate the same redundant links.



(a) GS^+ **(b)** Res

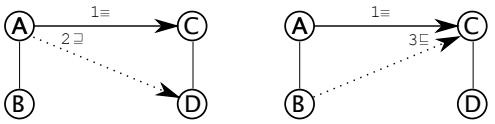
Figure 4. Minimization changing precision and recall

Consider the examples in the Figure 4. Suppose that all the displayed links are correct. Notice that links 2 and 3 follow from link 1. Suppose that our golden standard (Figure 4a), as it often happens with large datasets, is incomplete (it contains only the links 2 and 3, while link 1 is unknown) and thus we use precision formula (3) which returns an approximated value. Suppose that the matcher, being good enough, finds all displayed links (Figure 4b). By computing the precision and recall figures first on the original and then on the minimized versions of the mapping and the golden standard we obtain:

$$GS^+ = \{2,3\} \quad Res = \{1,2,3\}$$
$$Precision = 0.66 \quad Recall = 1 \tag{5}$$
$$Min(GS^+) = \{2,3\} \quad Min(Res) = \{1\}$$
$$Precision = 0 \quad Recall = 0 \tag{6}$$

Compare the normal situation (5) with (6) that shows the situation when minimized sets are used to calculate precision and recall figures. From this example we see that precision and recall figures computed on the minimized versions are far from the real values and are unreliable.

Our last observation is that using maximized sets gives no preference to redundant or non-redundant links and leads to more accurate results. In particular, recall figure better shows the amount of information actually found by the system. If we maximize the sets we also decrease the number of unknown links and therefore we obtain a more accurate result.



(a) GS^+ **(b)** Res

Figure 5. Maximization changing precision and recall

Consider now the example in Figure 5. The precision and recall figures are given in (7) for the original sets and in (8) for the maximized ones.

$$GS^+ = \{1,2\} \quad Res = \{1,3\}$$
$$Precision = 0.5 \quad Recall = 0.5 \tag{7}$$
$$Max(GS^+) = \{1,2,3\} \quad Max(Res) = \{1,2,3\}$$
$$Precision = 1 \quad Recall = 1 \tag{8}$$

Maximizing a golden standard can also reveal some unexpected problems and inconsistencies. For instance, we can discover that even if GS^+ and GS^- are disjoint, $Max(GS^+)$ and $Max(GS^-)$ are not, namely, $Max(GS^+) \cap Max(GS^-) \neq \emptyset$. During our experiments with the TaxME2 golden standard [14], we discovered that there are two links in the intersection of GS^+ and GS^- and 2187 in the intersection of their maximized versions.

We conducted several experiments to study the differences between precision and recall measures when comparing the minimized and maximized versions of the golden standards with the minimized and maximized versions of the mapping returned by S-Match [15].

We used three different golden standards [1] already used in several evaluations. The first two datasets come from OAEI; they describe publications, contain few nodes and corresponding golden standard is exhaustive. It only contains equivalence links. The second two come from the arts domain and the golden standard is crafted by experts manually. The third two datasets have been extracted from the Looksmart, Google and Yahoo! web directories. The golden standard is part of the TaxME2 and is extensively described in [14]. Unfortunately, all these golden standards suffer to a certain degree from

Table 1. Precision and Recall for minimized, normal, and maximized sets

Dataset pair	Precision, %			Recall, %		
	min	res	max	min	res	max
101/304	32.47	9.75	69.67	86.21	93.10	92.79
Topia/Icon	16.87	4.86	45.42	10.73	20.00	42.11
Source/Target	74.88	52.03	48.40	10.35	40.74	53.30

the problems described in the previous sections, thus the measures obtained must be considered as indicative.

Table 1 contains precision and recall figures calculated using standard precision and recall formulas (1) and (2). For the cases where no GS^- is provided, (3) is used instead of (1). In particular, these figures are the result of the comparison of the minimized mapping with the minimized golden standards (min), the original mapping with the original golden standards (res) and the maximized mapping with the maximized golden standards (max) respectively. As it can be noted from the measures obtained comparing the maximized versions with the original versions, the performance of the algorithm is on average better than expected. To summarize:

Recommendation 3. To obtain accurate measures it is fundamental to maximize both the golden standard and the matching result.

5 Conclusions and future work

In this paper, we proposed some recommendations to follow when building golden standards and to effectively use them to evaluate matching algorithms. Following these recommendations will make the evaluation and comparison of different algorithms more accurate.

We also discussed the issue of evaluating mappings with disjointness. In the current state of the art, no golden standard is available that explicitly provides true disjointness links. Moreover, disjointness is often confused with overlap. Thus, it is currently impossible to evaluate the performance of such algorithms.

In the future we will explore how the size of the golden standard influences the evaluation and how large should be the part covered by GS^+ and GS^- to be statistically significant as well as describe the methodology for evaluating rich mappings by supporting our recommendations with further experimental results.

REFERENCES

[1] Aliaksandr Autayeu, Vincenzo Maltese, and Pierre Andrews, 'Best practices for ontology matching tools evaluation', Technical report, University of Trento, DISI, (2009).

[2] Paolo Avesani, Fausto Giunchiglia, and Mikalai Yatskevich, 'A large scale taxonomy mapping evaluation', in *Proc. of ISWC*, pp. 67–81, (2005).

[3] Paolo Bouquet, Luciano Serafini, and Stefano Zanobini, 'Semantic coordination: A new approach and an application', in *Proc. of ISWC*, pp. 130–145, (2003).

[4] Jérôme David and Jérôme Euzenat, 'On fixing semantic alignment evaluation measures', in *Proc. of the Third International Workshop on Ontology Matching*, (2008).

[5] Jérôme David, Fabrice Guillet, and Henri Briand, 'Association rule ontology matching approach', *International Journal on Semantic Web and Information Systems*, **3**(2), 27–49, (2007).

[6] Hong Hai Do, Sergey Melnik, and Erhard Rahm, 'Comparison of schema matching evaluations', in *Proc. of the 2nd International Workshop on Web Databases*, pp. 221–237, (2002).

[7] Hong Hai Do and Erhard Rahm, 'Coma — a system for flexible combination of schema matching approaches', in *VLDB*, pp. 610–621, (2002).

[8] Marc Ehrig and Jérôme Euzenat, 'Relaxed precision and recall for ontology matching', in *Proc. of Integrating Ontologies Workshop*, (2005).

[9] Jérôme Euzenat, 'Semantic precision and recall for ontology alignment evaluation', in *Proc. of IJCAI*, pp. 348–353, (2007).

[10] Jérôme Euzenat and Pavel Shvaiko, *Ontology Matching*, Springer-Verlag, 2007.

[11] Fausto Giunchiglia, Vincenzo Maltese, and Aliaksandr Autayeu, 'Computing minimal mappings', Technical report, University of Trento, DISI, (2008).

[12] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich, 'Semantic schema matching', in *Proc. of CoopIS*, pp. 347–365, (2005).

[13] Fausto Giunchiglia, Dagobert Soergel, Vincenzo Maltese, and Alessandro Bertacco, 'Mapping large-scale knowledge organization systems', in *Proc. of ICSD*, (2009).

[14] Fausto Giunchiglia, Mikalai Yatskevich, Paolo Avesani, and Pavel Shvaiko, 'A large dataset for the evaluation of ontology matching systems', *The Knowledge Engineering Review Journal*, **24**, 137–157, (2008).

[15] Fausto Giunchiglia, Mikalai Yatskevich, and Pavel Shvaiko, 'Semantic matching: algorithms and implementation', in *Journal on Data Semantics, IX*, (2007).

[16] Fausto Giunchiglia and Ilya Zaihrayeu, 'Lightweight ontologies', in *Encyclopedia of Database Systems*, 1613–1619, (2009).

[17] Antoine Isaac, Shenghui Wang, Claus Zinn, Henk Matthezing, Lourens van der Meij, and Stefan Schlobach, 'Evaluating thesaurus alignments for semantic interoperability in the library domain', *IEEE Intelligent Systems*, **24**(2), 76–86, (2009).

[18] Siddharth Kaza and Hsinchun Chen, 'Evaluating ontology mapping techniques: An experiment in public safety information sharing', *Decision Support Systems*, **45**(4), 714–728, (2008).

[19] Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm, 'Generic schema matching with cupid', in *VLDB*, pp. 49–58, (2001).

[20] Christian Meilicke and Heiner Stuckenschmidt, 'Incoherence as a basis for measuring the quality of ontology mappings', in *Proc. of the 3rd International Workshop on Ontology Matching*, (2008).

[21] Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm, 'Similarity flooding: A versatile graph matching algorithm and its application to schema matching', in *Proc. of ICDE*, (2002).

[22] Natalya Fridman Noy and Mark A. Musen, 'Evaluating ontology-mapping tools: Requirements and experience', in *Proc. of OntoWeb-SIG3 Workshop*, pp. 1–14, (2002).

[23] Marta Sabou and Jorge Gracia, 'Spider: Bringing non-equivalence mappings to OAEI', in *Proc. of the Third International Workshop on Ontology Matching*, (2008).

[24] Willem Robert van Hage, Antoine Isaac, and Zharko Aleksovski, 'Sample evaluation of ontology-matching systems', in *Proc. of EON*, pp. 41–50, (2007).