Recommendations for Qualitative Ontology Matching Evaluations

Aliaksandr Autayeu, Vincenzo Maltese, and Pierre Andrews

DISI, University of Trento, Italy

Abstract. This paper suggests appropriate rules to set up ontology matching evaluations and for golden standard construction and use which can significantly improve the quality of the precision and recall measures.

We focus on the problem of evaluating ontology matching techniques [1] which find mappings with *equivalence*, *less general*, *more general* and *disjointness*, and on how to make the evaluation results fairer and more accurate.

The literature discusses the appropriateness and quality of the measures [2], but contains little about evaluation methodology [3]. Closer to us, [4] raises the issue of evaluating non-equivalence links.

Golden standards (GS) are fundamental for evaluating the precision and recall [2]. 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, in big datasets annotating all links is impractical and golden standards are often a sample of all node pairs, leading to approximate evaluations [5]. However, most current evaluation campaigns tend to use tiny ontologies, risking biased or poorly significant results.

> *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, results reliability depends on: (a) the portion of the pairs covered; (b) the ratio between GS^+ and GS^- sizes and (c) their quality (see last recommendation).

Most matching tools produce *equivalence*, some also produce *less general* and *more general* relations, but few output *disjointness* [6]. This must be taken into account to correctly compare evaluations. Usually, only the presence of a relation is evaluated, regardless the kind. Moreover, *disjointness* (two completely unrelated nodes) is often confused with *overlap* (two nodes whose intersection is not empty) and both are put in the GS^- [5]. This leads to imprecise results.

> *Recommendation 2.* When presenting evaluation results, specify whether and how the evaluation takes into account the semantic relations kind.

We use the notion of redundancy [7] to judge the quality of a golden standard. We use the **Min(mapping)** function to remove redundant links (producing the *minimized mapping*) and the **Max(mapping)** function to add all

redundant links (producing the *maximized mapping*). Following [7] and staying within lightweight ontologies [8] guarantees that the maximized set is always finite and thus precision and recall can always be computed. The table below presents the measures obtained in our experiments with SMatch on three different datasets (see [6] for details). Comparing the measures obtained with the maximized versions (max) with the measures obtained with the original versions (res), one can notice that the performance of the algorithm is on average better than expected. In [6] we explain why comparing the minimized versions is not meaninful and we conclude that:

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

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

Maximizing a golden standard can also reveal unexpected problems and inconsistencies. For instance, we discovered that in TaxME2 [5] $|GS^+ \cap GS^-| = 2$ and $|Max(GS^+) \cap Max(GS^-)| = 2187$. In future work 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^-, as well as describe methodology for evaluating rich mappings by supporting our recommendations with experimental results.

References

1. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. JoDS **4** (2005) 146–171
2. David, J., Euzenat, J.: On fixing semantic alignment evaluation measures. In: Proc. of the 3rd Ontology Matching Workshop. (2008)
3. Noy, N.F., Musen, M.A.: Evaluating ontology-mapping tools: Requirements and experience. In: Proc. of OntoWeb-SIG3 Workshop. (2002) 1–14
4. Sabou, M., Gracia, J.: Spider: Bringing non-equivalence mappings to OAEI. In: Proc. of the 3rd Ontology Matching Workshop. (2008)
5. Giunchiglia, F., Yatskevich, M., Avesani, P., Shvaiko, P.: A large dataset for the evaluation of ontology matching systems. KERJ **24** (2008) 137–157
6. Autayeu, A., Maltese, V., Andrews, P.: Best practices for ontology matching tools evaluation. Technical report, University of Trento, DISI (2009)
7. Giunchiglia, F., Maltese, V., Autayeu, A.: Computing minimal mappings. In: Proc. of the 4th Ontology Matching Workshop. (2009)
8. Giunchiglia, F., Marchese, M., Zaihrayeu, I.: Encoding classifications into lightweight ontologies. JoDS **8** (2007) 57–81