Extending an ontology alignment system with BioPortal: a preliminary analysis*

Xi Chen[1], Weiguo Xia[1], Ernesto Jiménez-Ruiz[2], Valerie Cross[1]

[1] Miami University, Oxford, OH 45056, United States
[2] University of Oxford, United Kingdom

1 Introduction

Ontology alignment (OA) systems developed over the past decade produced alignments by using lexical, structural and logical similarity measures between concepts in two different ontologies. To improve the OA process, string-based matchers were extended to look up synonyms for source and target concepts in background or external knowledge sources such as general purpose lexicons, for example, WordNet.[3] Other OA systems such as SAMBO [8] and ASMOV [6] applied this approach but with specialized background knowledge, i.e. the UMLS Metathesaurus,[4] for the anatomy track of the Ontology Alignment Evaluation Initiative[5] (OAEI). Then a composition-based approach was proposed to use background knowledge sources such as Uberon[6] and the Foundational Model of Anatomy[7] (FMA) as intermediate ontologies [5] for the anatomy track. Here source concepts and target concepts are first mapped to the intermediate background ontology. If source and target concepts map to an exact match in the intermediate ontology, a mapping can be made between them. Other OA systems also followed with a composition-based approach using Uberon [1, 2].

One issue on the use of background knowledge sources is determining the best knowledge source on which to use these various alignment techniques. Previous OA systems using specialized knowledge sources have pre-selected specific biomedical ontologies such as Uberon for the anatomy track.

As a coordinated community effort, BioPortal [3, 4] provides access to more than 370 biomedical ontologies, synonyms, and mappings between ontology entities via a set of REST services.[8] By tapping into this resource, an OA system has access to the full range of these ontologies, including Uberon and many of the ontologies integrated in the UMLS Metathesaurus. Since BioPortal has not been exploited in the context of the OAEI, this paper examines two practical uses of BioPortal as a generalized yet also specialized background knowledge source for the biomedical domain. We provide a preliminary investigation of the results of these two uses of BioPortal in the OAEI's anatomy track using the LogMap system [7].

* This research was financed by the Optique project with grant agreement FP7-318338

[3] http://wordnet.princeton.edu/
[4] http://www.nlm.nih.gov/research/umls
[5] http://oaei.ontologymatching.org
[6] http://obophenotype.github.io/anatomy/
[7] http://sig.biostr.washington.edu/projects/fm/AboutFM.html
[8] http://data.bioontology.org/documentation

Algorithm 1 Algorithm to assess border-line mappings using BIOPORTAL

Input: $m = \langle e_1, e_2 \rangle$: mapping to assess; τ_1,τ_2: thresholds; **Output:** true/false

Extract set of similar entities E_1 from BIOPORTAL for entity e_1
Extract set of similar entities E_2 from BIOPORTAL for entity e_2
if $E_1 \neq \emptyset$ and $E_2 \neq \emptyset$ **then**
    if $\mathsf{JaccardIndex}(\mathsf{E_1}, \mathsf{E_2}) > \tau_1$ **then**
        return true
    Extract mappings M_1 from BIOPORTAL for entities in E_1
    Extract mappings M_2 from BIOPORTAL for entities in E_2
    if $M_1 \neq \emptyset$ and $M_2 \neq \emptyset$ and $\mathsf{JaccardIndex}(\mathsf{M_1}, \mathsf{M_2}) > \tau_2$ **then**
        return true
return false

2 BIOPORTAL as an Oracle

Over the last few years, OA systems have made only minor improvements based on alignment performance measures of precision, recall, and F-score. This experience provides evidence that a performance upper bound is being reached using OA systems which are completely automatic. To increase their performance, some OA systems (e.g. LogMap) have included a semi-automatic matching approach which incorporates user interaction to assess borderline alignments (i.e. non "clear cut" cases with respect to their confidence values). For example, LogMap identifies 250 borderline mappings in the OAEI's anatomy track when its interactive mode is active.

The research presented in this paper investigates replacing the human expert with an automated expert or "oracle" that relies on specialized knowledge sources in the biomedical domain. BIOPORTAL provides access to different resources including a wide variety of ontologies, classes within ontologies and mappings between the classes of different ontologies. For example, BIOPORTAL allows to search for ontology classes whose labels have an exact match with a given term. The oracle can use this capability to assist in determining whether a borderline mapping produced by an OA system should be included in the final alignment output or not (i.e. increasing its confidence).

Algorithm 1 shows the implemented method to assess a given mapping m between entities e_1 and e_2 using BIOPORTAL as an oracle.

3 BIOPORTAL as a Mediating Ontology Provider

Mediating ontologies are typically pre-selected specifically for the OA task. For example the top systems in the OAEI's anatomy track used Uberon as (pre-selected) mediating ontology [5, 1, 2]. Limited research, however, has addressed the challenge of automatically selecting an appropriate mediating ontology as background knowledge [10, 9]. This research investigates using BIOPORTAL as a (dynamic) provider of mediating ontologies instead of relying on a few preselected ontologies.

Unlike [10] and [9], due to the large number of ontologies available in BIOPORTAL, we have followed a *fast-selection approach* to identify a suitable set of mediating

Algorithm 2 Algorithm to identify mediating ontologies from BIOPORTAL

Input: $\mathcal{O}_1$, $\mathcal{O}_2$: input ontologies; LM: a lexical matcher; N: stop condition
Output: Top-5 (candidate) mediating ontologies $\mathcal{MO}$

```
Compute exact mappings M between O1 and O2 using the lexical matcher LM
Extract representative entity labels S from M
for each label ∈ S
    Get ontologies from BIOPORTAL that contains an entity with label label (search call)
    Add to MO the ontologies that provides synonyms for label (record positive hits I)
    Record number of synonyms (II)
    Record ontology information: # of classes (III), depth (IV) and DL expressiveness (V)
    stop condition: if after N calls to BIOPORTAL MO did not change then stop iteration
return Top-5 ontologies from MO according to the number of positive hits and synonyms
```

Table 1: Top 5 mediating (BIOPORTAL) ontologies for the OAEI's anatomy track

#	Ontology	% pos. hits (I)	Avg. # syn. (II)	# classes (III)	Depth (IV)	DL exp. (V)
1	**SNOMED CT**	60%	5.1	401,200	28	$\mathcal{ALER}$
2	**UBERON**	63%	3.3	12,091	28	$\mathcal{SRIQ}$
3	**MeSH**	34%	5.0	242,262	16	$\mathcal{AL}$
4	**EFO**	16%	5.1	14,253	14	$\mathcal{SROIF}$
5	**CL (Cell Onto.)**	22%	3.3	5,534	19	$\mathcal{SH}$

ontologies from BIOPORTAL (see Algorithm 2[9]). The fast-selection approach identifies entity labels that appear in the input ontologies and searches to find ontologies in BIOPORTAL that include those labels and contain synonyms for them. The algorithm stops if the number of identified mediating ontologies does not change after a specified number N of (search) calls to BIOPORTAL or when there are no more labels to check.

Table 1 shows the identified top-5 mediating ontologies for the OAEI's anatomy track (with N=25 as stop condition). The ranking is based on the number of labels (i.e. search calls to BIOPORTAL) for which an ontology is able to provide synonyms (positive hits, **I**) and the average number of provided synonyms per positive hit (**II**). Additionally, information about the ontology is also given (**III–V**).

4 Preliminary evaluation

We have conducted a preliminary evaluation of the use of BIOPORTAL as a background knowledge provider (e.g. oracle and mediating ontology provider) in the OAEI's anatomy track and with LogMap as OA system. For this purpose, we have extended LogMap's matching process to *(i)* use Algorithm 1 as an oracle within its interactive mode (see Figure 3 in [7]); and *(ii)* use a mediating ontology $\mathcal{MO}$ as in Algorithm 3.

The results[10] are summarized in Table 2. Last column shows the original scores produced by LogMap (without BIOPORTAL). As expected, the best results in terms of

[9] In the close future, we plan to combine this algorithm with the ontotology recommender provided by BIOPORTAL: `https://bioportal.bioontology.org/recommender`

[10] SNOMED and MeSH have been discarded as mediating ontologies. SNOMED is not available to download, and we were unable to download MeSH due to a time-out given by BIOPORTAL.

Algorithm 3 Use of a mediating ontology with LogMap

Input: $\mathcal{O}_1$, $\mathcal{O}_2$: input ontologies; $\mathcal{MO}$: mediating ontology; **Output:** $\mathcal{M}$: output mappings;

1: $\mathcal{M}_1 := \mathsf{LogMap}(\mathcal{O}_1, \mathcal{MO})$
2: $\mathcal{M}_2 := \mathsf{LogMap}(\mathcal{MO}, \mathcal{O}_2)$
3: $\mathcal{M}_C := \mathsf{ComposeMappings}(\mathcal{M}_1, \mathcal{M}_2)$
4: $\mathcal{M} := \mathsf{LogMap}(\mathcal{O}_1, \mathcal{O}_2, \mathcal{M}_C)$
5: **return** $\mathcal{M}$

Table 2: Results of LogMap with/without BIOPORTAL as background knowledge

Score \ Mode	LogMap - BIOPORTAL				LogMap
	Oracle	$\mathcal{MO}_{\text{Uberon}}$	$\mathcal{MO}_{\text{CL}}$	$\mathcal{MO}_{\text{EFO}}$	
Precision	**0.915**	0.899	0.907	0.914	0.913
Recall	0.846	**0.927**	0.867	0.846	0.846
F-score	0.879	**0.913**	0.886	0.879	0.878

F-score has been obtained using Uberon as mediating ontology. Using CL as mediator also improves the results with respect to those obtained by LogMap, although the improvement does not have an impact as big as with Uberon. There is not significant improvement using EFO as mediating ontology. Using BIOPORTAL as an oracle leads to a small increase in precision, but recall remains the same.

This preliminary evaluation has shown the potential of using BIOPORTAL as background knowledge. In the close future we plan to conduct an extensive evaluation involving more challenging datasets (e.g. OAEI's largebio track) and other OA systems, and combining several mediating ontologies.

References

1. Cruz, I.F., Stroe, C., Caimi, F., Fabiani, A., Pesquita, C., Couto, F.M., Palmonari, M.: Using AgreementMaker to Align Ontologies for OAEI 2011. In: 6th OM Workshop (2011)
2. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F., Couto, F.M.: The AgreementMakerLight Ontology Matching System. In: OTM Conferences. pp. 527–541 (2013)
3. Fridman Noy, N., Shah, N.H., Whetzel, P.L., Dai, B., et al.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research 37, 170–173 (2009)
4. Ghazvinian, A., Noy, N.F., Jonquet, C., Shah, N.H., Musen, M.A.: What four million mappings can tell you about two hundred ontologies. In: Int'l Sem. Web Conf. (ISWC) (2009)
5. Gross, A., Hartung, M., Kirsten, T., Rahm, E.: Mapping composition for matching large life science ontologies. In: 2nd International Conference on Biomedical Ontology (ICBO) (2011)
6. Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology Matching with Semantic Verification. Journal of Web Semantics 7(3), 235–251 (2009)
7. Jiménez-Ruiz, E., Cuenca Grau, B., Zhou, Y., Horrocks, I.: Large-scale interactive ontology matching: Algorithms and implementation. In: European Conf. on Art. Int. (ECAI) (2012)
8. Lambrix, P., Tan, H.: A System for Aligning and Merging Biomedical Ontologies. Journal of Web Semantics 4(3), 196–206 (2006)
9. Quix, C., Roy, P., Kensche, D.: Automatic selection of background knowledge for ontology matching. In: Proc. of the Int'l Workshop on Sem. Web Inf. Management (2011)
10. Sabou, M., d'Aquin, M., Motta, E.: Exploring the Semantic Web as Background Knowledge for Ontology Matching. J. Data Semantics 11, 156–190 (2008)