Compound Matching of Biomedical Ontologies

Daniela Oliveira* and Catia Pesquita

LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Campo Grande
1749-016, Portugal

ABSTRACT

Biomedical ontologies are particularly successful in the uniformization of the life sciences domain and ontology matching systems are useful to discover relationships between concepts of two different ontologies. However, that is also a limitation as there is a growing interest in discovering more complex kinds of mappings and existing techniques are limited to matching two ontologies. Therefore, producing 'compound' alignments, which match more than two ontologies, could be potentially useful to support a next generation of semantic technologies.

In this paper, we present a novel algorithm that produces compound matches between three different ontologies and its performance is evaluated against seven automatically inferred reference alignments from the biomedical domain. We analyze all alignments manually to verify the results and propose a new way to complete the logical definitions of OBO cross-products.

1 INTRODUCTION

Biomedical ontologies typically contain a high number of classes and many times cover the same field or related fields, which hinders their interoperability. One approach to address this problem is the use of matching systems which are capable of establishing meaningful connections between ontologies.

Still, most ontology matching systems produce equivalence mappings between classes or properties in two ontologies. However, in a complex domain such as biomedicine, where several ontologies describe different but related aspects of biomedical phenomena, it may be advantageous to create mappings by combining entities from more than two ontologies. We argue that it would be useful for the developers of ontology alignment systems to develop new techniques and tools for identifying 'compound matches', i.e. matches between class or property expressions involving more than two ontologies. To the best of our knowledge, there are currently no ontology matching systems capable of generating such mappings.

The purpose of this work is to develop novel algorithms which can be used for the efficient and effective creation of alignments between a class A of one ontology with an expression relating classes B and C of two other ontologies, constituting a ternary relationship.

2 METHODS

We consider that a ternary compound alignment is a set of correspondences (mappings) between classes from a source ontology O$_s$ and class expressions obtained by combining two other classes each belonging to a different target ontology O$_{t1}$ and O$_{t2}$ (see Figure 1). This means that we define a ternary compound mapping as a tuple <X, Y, Z, R, M>, where X, Y and Z are classes from three distinct ontologies, R is a relation established between Y and Z to generate a class expression that is mapped to X via a mapping relation M. Here, we consider the ontology to which X belongs to be the source ontology, and the ontologies that define Y and Z to be the target ontology 1 and 2, respectively. In this particular case the relation R is always an intersection (regardless of any qualifier) and the mapping M an equivalence.

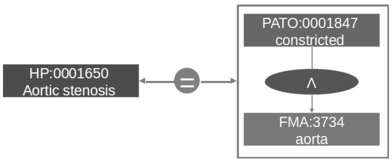

Fig. 1. Example of a possible ternary compound match.

2.1 Implementation

We developed a novel algorithm to establish compound mappings integrated into the AgreementMakerLight (AML) (Faria *et al.*, 2014) ontology matching system[1]. Our algorithm exploits AML's *Word Lexicon*, the set of all words in an ontology's vocabulary to which are assigned an evidence content (EC), reflecting the usage of the word within the ontology.

In a first step, we perform a pairwise mapping of the labels of O$_s$ with the labels of O$_{t1}$, by the ratio of the sum of the EC of the words shared by the source label (l$_s$) and the target 1 label (l$_{t1}$), and the sum of the EC of the words in l$_{t1}$.

$$sim(l_s, l_{t1}) = \frac{\sum EC(word \in (l_s \cap l_{t1}))}{\sum EC(word \in l_{t1})} \tag{1}$$

We filter out all mappings with similarity below a given threshold. In a second step, for each mapping found in step 1, we remove from the source labels all the words that have already been matched (l_{s*}). Taking as an example the mapping in Figure 1, after matching HP and FMA, which would capture the mapping for 'aorta', the HP's class label would be reduced to 'stenosis'.
In a third step, for each mapping, we perform a pairwise comparison of the reduced source labels with target 2 labels. However, here the ratio divisor corresponds to the sum of EC of the words in the label with more words, to ensure the longest possible match.

$$sim(l_s, l_{t2}) = \frac{\sum EC(word \in (l_{s*} \cap l_{t2}))}{\sum EC(word \in longest(l_s, l_{t2}))} \tag{2}$$

In a fourth step, the final similarity between the matched labels is computed as the average between the similarities computed in steps 1 and 3. Label mappings below the second threshold are filtered out. Finally, the algorithm has a greedy selection step, which selects the

*To whom correspondence should be addressed: doliveira@lasige.di.fc.ul.pt

[1] Available at: https://github.com/AgreementMakerLight/AML-Compound

mapping with the highest similarity, amongst the source classes with more than one mapping.

2.2 Evaluation

To evaluate our strategy we used a set of seven reference alignments (Pesquita *et al.*, 2014) automatically created by inferring compound mappings from cross-products (Mungall *et al.*, 2011) of the logical definitions in OBO ontologies (Smith *et al.*, 2007). For this, we computed precision, recall and f-measure. We also performed a manual evaluation of the results, where we classified mappings into three possible categories: 'Correct', where the mapping is deemed correct and the source class has no mapping in the reference alignment; 'Conflict', where the mapping is deemed correct but the source class has a different mapping in the reference alignment; and 'Incorrect', where the mapping is deemed incorrect. We applied this to all mappings created by using 0.5 as a threshold for step 1 and 0.9 for step 2.

3 RESULTS

Table 1 presents some statistics about the alignments obtained. Preliminary results using this evaluation approach present low F-Measure, with a higher precision, which fluctuates between 67.9 and 11.6 and recalls that always fall below the 50% mark.

	Precision	Recall	F-Measure
MP-CL-PATO	52.6 %	20.8 %	29.8 %
MP-GO-PATO	67.9 %	47.2 %	55.7 %
MP-NBO-PATO	47.3 %	30.1 %	36.8 %
MP-UBERON-PATO	64.7 %	19.4 %	29.9 %
WBP-GO-PATO	11.6 %	7.7 %	9.2 %
HP-FMA-PATO	21.2 %	12.4 %	15.6 %

Table 1. Evaluation results from the comparison with the automated reference alignments

	Correct	Conflict	Incorrect
MP-CL-PATO	63.71 %	34.60 %	1.69 %
MP-GO-PATO	92.16 %	6.97 %	0.87 %
MP-NBO-PATO	72.46 %	26.09 %	1.45 %
MP-UBERON-PATO	91.33 %	7.96 %	0.70 %
WBP-GO-PATO	88.55 %	7.49 %	3.96 %
HP-FMA-PATO	77.82 %	15.56 %	6.61 %

Table 2. Manual evaluation of results.

The manual inspection of the mappings (Table 2) revealed that the algorithm is finding mostly correct mappings, with the lowest percentage belonging to the MP-CL-PATO compound alignment, which had the highest number of conflicting mappings.

4 DISCUSSION

One challenge in computing compound alignments is the memory requirements involved in the process. If matching two large biomedical ontologies is already a challenge for many ontology matching systems, handling three ontologies in a compound alignment scenario is even more demanding. Our algorithm reduces the search-space by using the two-step matching approach, which both reduces the time and memory requirements [2].

Although our algorithm's performance against the reference alignments is low (Table 1), the manual evaluations of the mappings reveals a very low proportion of incorrect mappings, so we investigated how these new mappings could impact the logical definitions of the source ontology. The results presented in Table 3 indicate that the logical definitions of the three source ontologies could be expanded with more than 800 new logical definitions.

Ontology	New Mappings	OBO classes	% of Growth
MP	422	7694	5.48
WBP	182	957	19.02
HP	259	14059	1.84

Table 3. Influence of the new mappings on the source ontology.

We can conclude that our approach is capable of producing good precision (Table 2 shows an average of 81% of the matches are correct), and is able to find many correct mappings that are not in the reference alignment. However, it struggles with capturing many of the mappings in the references, which is mainly due to our algorithm's inability to distinguish between similar PATO class (e.g., PATO:0000470: 'present in greater numbers in organism' *vs.* PATO:0002002: 'has extra parts of type'), or the use of synonyms not defined in any of the ontologies.

5 CONCLUSION

We have presented, to the best of our knowledge, the first algorithm for compound matching of ontologies. It is particularly suited for biomedical ontologies, given its ability to handle large ontologies and the need in this domain to reveal more complex relations between them. Our preliminary experiments have shown that, despite the challenges in handling an increased matching space and the inherently more difficult-to-compute ternary mapping, our algorithm is able to produce good precision mappings. Moreover, we posit that it could also be used as a first step in adding new logical definitions to ontologies, since we were able to find several correct mappings that were not in the reference alignments..

ACKNOWLEDGEMENTS

The authors are grateful to Daniel Faria for his technical support. This work was supported by FCT through funding of LaSIGE Research Unit, ref.UID/CEC/00408/2013

REFERENCES

Faria, D., Pesquita, C., Santos, E., Cruz, I. F., and Couto, F. M. (2014). AgreementMakerLight: a scalable automated ontology matching system. *10th International Conference on Data Integration in the Life Sciences 2014 (DILS)*, page 29.

Mungall, C. J., Bada, M., Berardini, T. Z., Deegan, J., Ireland, A., Harris, M. A., Hill, D. P., and Lomax, J. (2011). Cross-product extensions of the Gene Ontology. *Journal of Biomedical Informatics*, **44**(1), 80 – 86. Ontologies for Clinical and Translational Research.

Pesquita, C., Cheatham, M., Faria, D., Barros, J., Santos, E., and Couto, F. M. (2014). Building reference alignments for compound matching of multiple ontologies using OBO cross-products. In *Ontology Matching Workshop at ISWC 2014*.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., *et al.* (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, **25**(11), 1251–1255.

[2] The largest alignment takes less than 15 minutes with an Intel® Core™i7-2600 CPU 3.40GHz x 8 processor and 16GB memory.