XMap++ : Results for OAEI 2014

Warith Eddine Djeddi and Mohamed Tarek Khadir

LabGED, Computer Science Department, University Badji Mokhtar, Annaba, Algeria
{djeddi, khadir}@labged.net

Abstract. In this paper, we present the results obtained by our ontology matching system XMap++ within the OAEI 2014 campaign. XMap++ is a scalable ontology alignment tools capable of matching large scale ontology. This is our second participation in the OAEI, and we can see an overall improvement on nearly every task.

1 State, purpose, general statement

XMap (eXtensible Mapping) is an ontology alignment tool for the alignment of OWL entities (i.e., classes, object properties and data properties). XMap++ approach uses different similarity measures of different categories such as string, linguistic, and structural based similarity measures to understand ontologies semantics. A weights vector must, therefore, be assigned to these similarity measures, if a more accurate and meaningful alignment result is favored. Combining multiple measures into a single similarity metric has been solved using weights determined by intelligent strategies [3].

The major drawback of our two previous versions XMapGen and XMapSig [2], despite the fact that they achieved fair results and the aim of their development is to deliver a stable version, the time performance was very low time, especially for the Large Biomedical Ontologies tracks, inability to recognize multiple labels to a single entity as synonyms and inability to recognize labels translated in different languages (e.g Chinese, Czech, Dutch, French, German). After carefully studying this issue, we realize that our algorithm needs more assessment in its performance. This inspires us to consider new strategies in the new version of XMap++ 2014, such as : 1) Using cosine similarity as a string similarity methods to compare the concepts textual descriptions associated with the nodes (labels, names, identity, etc) of each ontology; 2) Involving particular parallel matching on multiple cores or machines for dealing with the scalability issue on ontology matching; 3) Translating labels with different languages using Bing Translator (not use any services which require payment); 4) Interfacing with the Wordnet electronic dictionary using Java Wordnet Interface (JWI) as a Java library. Meanwhile, XMap++ loads WordNet dictionary fully into memory to gain time when it aligns large-scale ontologies. Consequently, the new version XMap++ 2014 has improved both the matching quality and time performance in large scale ontology matching tasks.

1.1 Specific techniques used

The workflow and the main components of the system can be seen in the Fig. 1. The XMap++ consists of the following components:

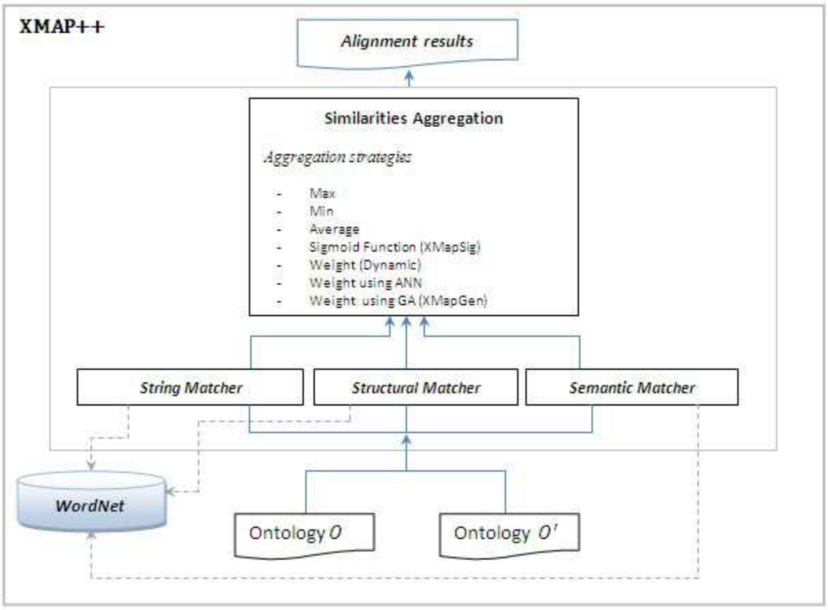

Fig. 1. Sketch of Architecture for XMAP++.

1. Matching inputs are two ontologies, source O and target $O^{'}$ parsed by an Ontology Parser component;
2. The ***String Matcher*** based on linguistic matching compares the textual descriptions of the concepts associated with the nodes (labels, names) of each ontology;
3. The ***Linguistic matcher*** jointly aims at identifying words in the input strings, relaying on WordNet [7]. These matching techniques may provide incorrect match candidates, structural matching is used to correcting such match candidates based on their structural context. In order to deal with lexical ambiguity, we introduce the notion of the *scope* belonging to a concept which represents the context where it is placed [1]. The value of linguistic methods is added to the linguistic matcher or the structure matcher in order to enhance the semantic ambiguity during the comparison process of entity names;
4. The ***structural matcher*** aligns nodes based on their adjacency relationships. The relationships (e.g., *subClassOf* and *is-a*) that are frequently used in the ontology serve, at one hand, as the foundation of the structural matching;
5. The three matchers perform similarity computation in which each entity of the source ontology is compared with all the entities of the target ontology, thus producing three similarity matrices, which contain a value for each pair of entities. After that, an aggregation operator is used to combine multiple similarity matrices computed by different matchers to a single aggregated similarity matrix. We refer to [3] for more detail about the pruning and splitting techniques on data matrices for two couple of entities;

6. XMap++ uses three types of aggregation operator; these strategies are *aggregation*, *selection* and *combination* [3];
7. Finally, these values are filtered using a selection according to a defined threshold and the desired cardinality. In our algorithm, we adopt the *1-1* cardinality to find the optimal solution in polynomial time.

2 Results

In this section, we present the evaluation results obtained by running XMap++ with SEALS client with *Benchmark*, *Anatomy*, *Conference*, *Multifarm*, *Library* and *Large Biomedical Ontologies* tracks. Adding to that, we present the results of the test *Ontology Alignment for Query Answering* which not follow the classical ontology alignment evaluation on the SEALS platform.

2.1 Benchmark

XMap++ performs very well in terms of Precision (1.0) while a low recall (0.4) in the Benchmark track. Those low values are explained by the fact that ontological entities with scrambled labels, lexical similarity becomes ineffective. Whereas for the others two test suites our algorithm performed worse in term of F-Measure because our system does not handle ontology instances. Table 1 summarises the average results obtained by XMap++.

Table 1. Results for Benchmark track.

Test	P	R	F
biblio	1.0	0.40	0.57
cose	1.0	0.17	0.28
dog	1.0	0.20	0.32

2.2 Anatomy

The Anatomy track consists of finding an alignment between the Adult Mouse Anatomy (2744 classes) and a part of the NCI Thesaurus (3304 classes) describing the human anatomy. XMap++ achieves a good F-Measure value of ≈89% in an adequate amount of time (22 sec.) (see Table 2). In terms of F-Measure/runtime, XMap++ ranked 3rd among the 10 tools participated in this track.

2.3 Conference

The Conference track uses a collection of 16 ontologies from the domain of academic conferences. Most ontologies were equipped with OWL DL axioms of various kinds; this opens a useful way to test our semantic matchers. The match quality was evaluated

Table 2. Results for Anatomy track.

System	Precision	F-Measure	Recall	Time(s)
XMap++	0.940	0.893	0.850	22

against the original (ra1) as well as entailed reference alignment (ra2). As the Table 3 shows, for both evaluations we achieved F-Measure values better than the two Baselines results (edna, StringEquiv).

Table 3. Results for Conference track.

System	RA1 Reference			RA2 Reference		
	P	R	F	P	R	F
XMap++	0.87	0.49	0.63	0.82	0.44	0.57

2.4 Multifarm

This track is based on the translation of the OntoFarm collection of ontologies into 9 different languages. XMap ++'s results are showed in the Table 4.

Table 4. Results for Multifarm track.

System	Different ontologies			Same ontologies		
	P	F	R	P	F	R
XMap++	0.31	0.35	0.43	0.76	0.50	0.40

2.5 Library

The library track involves the matching of the STW thesaurus (6,575 classes) and the Soz thesaurus (8,376 classes). Both of these thesauri provide vocabulary for economic and social sciences. The results are depicted in table 5; our tools achieved a good recall of $\approx$88%, and the precision was low (50%). XMap++ requires $\approx$ 3 hr and 30 min, it is mainly due to the long times required for looking up concepts in Bing Translator when it attempts to translate all the German labels to English labels.

Table 5. Results for Library track.

System	Precision	Recall	F-Measure	Time(s)
XMap++	0.508	0.885	0.646	12652

2.6 Large biomedical ontologies

This track consists of finding alignments between the Foundational Model of Anatomy (FMA), SNOMED CT, and the National Cancer Institute Thesaurus (NCI). There are 6 sub-tasks corresponding to different sizes of input ontologies (small fragment and whole ontology for FMA and NCI and small and large fragments for SNOMED CT). The results obtained by XMap++ are depicted on Table 6. In general we can conclude that Xmap++ achieved a good precision and fair recall value. The fair recall value can be explained by the fact that WordNet does not contain definitions of highly technical medical terms, resulting in the system being unable to match entities that are not located in the WordNet database. Using a different linguistic ontology should alleviate this problem, or ideally the system should automatically select the most appropriate linguistic ontology for this task.

Table 6. Results for the Large BioMed track.

Test set	Precision	Recall	F-Measure	Time(s)
Small FMA-NCI	0.932	0.848	0.888	17
Whole FMA-NCI	0.835	0.745	0.787	144
Small FMA-SNOMED	0.858	0.737	0.793	35
Whole FMA- Large SNOMED	0.558	0.633	0.593	390
Small SNOMED-NCI	0.849	0.665	0.746	182
Whole NCI- Large SNOMED	0.843	0.584	0.690	490

2.7 Ontology Alignment for Query Answering

The objective of this test is to verify the ability of the generated alignments to answer a set of queries in an ontology-based data access scenario where several ontologies exist. The table 7 shows the F-measure results for the whole set of queries. XMap++ was one of the four matchers whose alignments allowed to answer all the queries of the evaluation.

Table 7. Results for Ontology Alignment for Query Answering.

System	RA1 Reference			RAR1 Reference		
	P	R	F	P	R	F
XMap++	0.556	0.487	0.505	0.554	0.487	0.505

3 General comments

3.1 Comments on the results

This is the second time that we participate in the OAEI campaign. While we participated with two configurations of our system to the 2013 edition of the campaign, respectively with XMapGen and XMapSig, this year a unique version has been submitted. Several changes have been introduced. The official results of OAEI 2014 show that XMap++ is competitive with other well-known ontology matching systems in all OAEI tracks, especially in Library track it got the highest recall of all attended systems. The current version of XMap++ has shown a significant improvement both in terms of matching quality and runtime. Additionally, to tackle the large ontology matching problem we improved the runtime of the algorithm using a divide-and-conquer approach that can partition the execution of the matchers into small threads was improved and joins their results after each similarity calculation.

3.2 Discussions on the way to improve the proposed system

Some probable approaches to improving our tools are listed as follows:

1. Take comments and Instance information of ontology into account, especially when the name of a concept is meaningless;
2. Using the UMLS Meta-thesaurus to have high recall when aligning ontologies from the biomedical science domain;
3. Pre-compiling a local dictionary in order to avoid multiple accesses to the Microsoft Translator within the matching process.

3.3 Comments on the OAEI 2013 procedure

As a second participation, we found the OAEI procedure very convenient and the organizers very supportive. The use of Seals allows objective assessments. The OAEI test cases are various, and this leads to comparison on different levels of difficulty, which is very interesting. We found that SEALS platform is a very valuable tool to compare the performance of our system with the others.

4 Conclusion

We have briefly described our fully automate ontology matching system XMap++ and presented the results achieved during the 2014 edition of the OAEI campaign. The obtained results showed that XMap++ is able to efficiently and effectively match ontologies of different size. In future we want to participate in more tracks. Our ontology matching system presents some limitations. We intend to use the UMLS resource for better discarding incorrect mappings for life sciences related ontologies.

References

1. Djeddi, W., Khadir, M. T.: A Novel Approach Using Context-Based Measure for Matching Large Scale Ontologies. In Proceedings of 16th International Conference on Data Warehousing and Knowledge Discovery (DAWAK 2014), September 2-4, pp. 320–331. Springer, Munich, Germany (2014)
2. Djeddi, W., Khadir, M. T.: XMapGen and XMapSiG results for OAEI 2013. In Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), October 21, pp. 203–210. CEUR-WS.org, Sydney, Australia (2013)
3. Djeddi, W., Khadir, M.T.: Ontology alignment using artificial neural network for large-scale ontologies. In the International Journal of Metadata, Semantics and Ontologies (IJMSO), Vol.8, No.1, pp.75-92 (2013)
4. Djeddi, W., Khadir, M.T.: Introducing artificial neural network in ontologies alignment process. In the Journal Control and Cybernetics, Vol. 41, No. 4, pp.743-759 (2012)
5. Djeddi, W., Khadir, M.T. : A dynamic multistrategy ontology alignment framework based on semantic relationships using WordNet. In Proc of the 3rd International Conference on Computer Science and its Applications (CIIA′11), $13-15$ December, Saida, Algeria, pp.149-154 (2011)
6. Djeddi, W., Khadir, M.T.: XMAP: a novel structural approach for alignment of OWL-full ontologies. In Proc. of the International Conference on Machine and Web Intelligence (ICMWI), pp.347-352 (2010)
7. Fellbaum, C. : WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA (1998)
8. Gross, A., Hartung, M., Kirsten, T. and Rahm, E. : On matching large life science ontologies in parallel. In , in Lambrix, P. and Kemp, G.J.L. (Eds), DILS, Springer, pp.35-49 (2010)