KEPLER at OAEI 2018

Marouen KACHROUDI[1], Gayo DIALLO[2], and Sadok BEN YAHIA[1]

[1] Université de Tunis El Manar, Faculté des Sciences de Tunis
Informatique Programmation Algorithmique et Heuristique
LR11ES14, 2092, Tunis, Tunisie
{marouen.kachroudi,sadok.benyahia}@fst.rnu.tn
[2] BPH Center - INSERM U1219, Team ERIAS & LaBRI UMR5800,
Univ. Bordeaux, F-33000, France
gayo.diallo@u-bordeaux.fr

Abstract. This paper presents and discusses the results of the KEPLER system for the 2018 edition of the Ontology Alignment Evaluation Initiative (OAEI 2018). The implemented approach is based on the exploitation of three different strategies including Information Retrieval (IR) inspired algorithm for terminological based alignment computation. For scaling up, KEPLER implements a partitioning approach, while for the management of multilingualism, KEPLER develops a well-defined strategy based on the use of a translator and structural alignment computation. This is the second year of participation and the results are encouraging.

1 Presentation of the system

A substantial growth of the semantic Web users create and update knowledge resources all over the world using various conceptualizations. These knowledge resources are used for annotating available online data. This process is nowadays being accelerated due to few initiatives which encourage to make data available in a comprehensive way for agents [1]. However, as they are annotated by different conceptual schemes, an effort is needed to make them interoperable. As of a solution, ontology alignment process is applied in order to identify bridges between the heterogeneous knowledge resources (ontologies, structured vocabularies, etc.) which play the role of semantic background for the available data. This process facilitates the share and reuse of these resources [2].

KEPLER is an ontology alignment system which deals with the key challenges related to heterogeneous ontologies on the semantic Web. It is grounded from previous approaches [3–6] and relies on several alignment strategies summarized in the following sections. It is designed to discover alignments for both common size and large scale ontologies as well as computing alignments in a multilingual context.

1.1 State, purpose, general statement

KEPLER exploits, besides classic techniques [7], an external resource, *i.e.*, a translator in order to deal with multilingualism.

1.2 Specific techniques used

The main idea of KEPLER is to exploit the expressiveness of the OWL language to detect and compute the similarity between entities of two given ontologies through six complementary modules as presented in Figure 1.

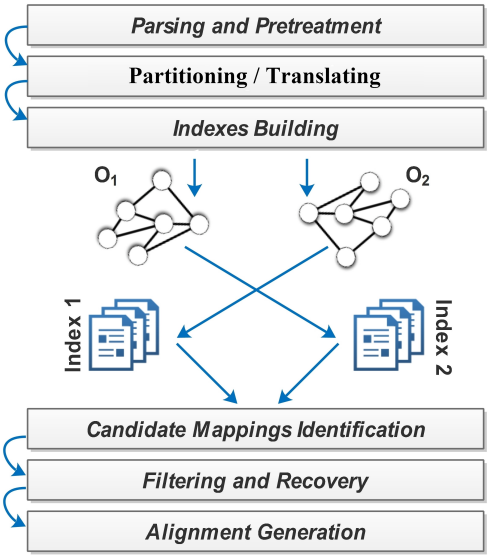

Fig. 1. KEPLER workflow.

Entities are described using OWL primitives with their semantics. An ontology is seen as a semantic graph where entities are nodes connected by links (the predicates). These links have specified semantics. The alignment workflow is detailed as follows.

Parsing and pretreatment: this module extracts the ontological entities initially represented by a primary form of lists. In other words, at the parsing stage, the main goal is to transform an OWL ontology in a well defined structure that preserves and highlight all the information contained in processed ontology. It has a significant impact on the results of the similarity computation thereafter. The result is a set of entities names and their associated descriptions.

Partitioning: KEPLER follows a divide and conquer strategy. Therefore, this module aims at splitting ontologies into smaller parts to support the alignment task [8]. Consequently, partitioning a set $\mathcal{B}(\mathcal{C})$ is to find subsets $\mathcal{B}_1$, $\mathcal{B}_2$,..., $\mathcal{B}_n$, encompassing semantically close elements bound by a relevant set of relationships, *i.e.*, $\mathcal{O} = \bigcup\{\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_n\}$, where $\mathcal{B}_i$ is an ontological block, and n is the resulting number of extracted blocks. Hence, we can define an ontological portion as a reduced ontology that could be extracted from another larger one by splitting up the latter according to its constituents : structures and semantics. One way to obtain such a partitioning is to maximize the relationships inside a block while minimizing the relationship between the

blocks themselves. The resulting partitioning quality can be evaluated using different criteria:

- *The size of the generated blocks*: that must have a reasonable size, *i.e.*, a number of elements that could be handled by an alignment tool;
- *The number of the generated blocks*: this number should be as small as possible to limit the number of block pairs to be aligned latter;
- *The compactness degree of a block*: a block is said to be substantially compact if relations (lexical and structural ones) are stronger inside the block and lower outside it.

Translation : in order to deal with multilingualism, two alternatives are followed: *i*) either considering one of the languages of the input ontologies as a pivot, therefore translating the second one to this chosen pivot; *ii*) choosing a pivot language and translate the inputs ontologies to this pivot. Further to these alternatives, an external resource, *i.e.*, WordNet[3] is used. Therefore the pivot language used by KEPLER is the English language. The translation process is performed usinig the Microsoft Bing [4] translator.

Indexing : one of the issue in Ontology Alignment is the cost of computing the similarity between all the entities of the input ontologies. To deal with this issue, the indexing strategy is one of the novelties of our approach. It consists in reducing the search space through the use of techniques borrowed from the IR domain. An effective search strategy is implemented on top of the built indexes of the two input ontologies. To enable faster searching, the driving idea that was previously used in the ServOMap system [9] is to perform the analysis of the ontologies in advance and store it in an optimized format for the search.

Candidate Mappings Identification : the role of this module is to find the entities in common between the indexes. Once the indexes are set up, the querying step is activated. To do so, the querying strategy implemented satisfies both the terminology search and semantic aspects at once. Indeed, the task is querying documents in a vector space that contains a set of ontological entities and their synonyms obtained via WordNet for each Ontology. It is worthy to mention that indexes querying is done in both senses (each ontology plays successively the role of querying component).

Filtering and Recovery: the filtering module consists of two complementary submodules, each one is responsible of a specific task in order to refine the set of primarily identified candidates mappings. At this stage, once the list of candidates is ready, the alignment method uses a first filter. This filter eliminates the redundancy between these candidates by eliminating possible duplicates. In addition, there is always the concern about *false positives*. The second filter eliminates *false positives* candidates. This filter is applied to what is called *partially* redundant entities. An entity is considered as *partially* redundant if it belongs to two different mappings. Being given three ontological

[3] https://wordnet.princeton.edu/
[4] https://www.bing.com/translator

entities e_1, e_2 and e_3, if on the one hand, e_1 is aligned to e_2, and secondly, e_1 is aligned to e_3, this last alignment is qualified as doubtful. As the KEPLER system generates $(1:1)$ mappings, an entity cannot belongs to several mappings. Therefore, given the topology of two suspicious entities (e_3 neighbors with e_1 neighbors, e_2 neighbors with e_1 neighbors) with respect to the redundant entity e_1, the idea is to retain the couple having the highest topological proximity value. All candidates are subject to this filter before to generate the final alignment.

Alignment Generation : The result of the alignment process provides a set of mappings, which are serialized in the RDF format.

2 Results

In this section, we present the results obtained by KEPLER system for the OAEI 2018 edition.

2.1 Anatomy track

This track consists in two real world ontologies to be matched, the source ontology describing the Adult Mouse Anatomy (with 2744 classes) and the target ontology is the NCI Thesaurus describing the Human Anatomy (with 3304 classes). For this track, KEPLER succeeded to extract 74% of correct mappings with a precision of 95% and recall of 74%. KEPLER handled easyly the input ontologies of this track thanks to the partionning module $Ontopart$ [10, 8]

2.2 Conference track

The conference track consists of 15 ontologies from the conference organization domain and each ontology must be matched against every other ontologies. The dataset describes the domain of organizing conferences from different perspectives. Precision values for to evaluation settings are respectively 76% and 58%. Recall values are 48% and 68%.

2.3 Multifarm track

The Multifram dataset is composed of a subset of the Conference track, translated in nine different languages (*i.e.*, Chinese, Czech, Dutch, French, German, Portuguese, Russian, Spanish and Arabic). With a special focus on multilingualism, it is possible to evaluate and compare the performance of alignment approaches through these test cases. Based on several previous contributions [11–16], the designed main goal of the MultiFarm track is to evaluate the ability of the alignment systems to deal with multilingual ontologies. It serves the purpose of evaluating the strength and weakness of a given system across languages. In the *different ontologies* setting, KEPLER is ranked second with a recall value of 0.21 and a precision value of 0.40. Whereas in the *same ontologies* setting, it lasted at the first place with a recall value of 0.36, and a precision value of 0.85.

2.4 Complex track

KEPLER succeeds in the best case, to obtain 27% of recall and a precision of 100%.

2.5 Large Biomedical Ontologies track

In the scalability register, this track consists in finding alignments between the Foundational Model of Anatomy (FMA), SNOMED CT, and the National Cancer Institute Thesaurus (NCI). These ontologies are semantically rich and contain tens of thousands of classes. The Large BioMed Track consists of three matching problems, *i.e.*, (1) FMA-NCI matching problem, (2) FMA-SNOMED matching problem and (3) SNOMED-NCI matching problem. KEPLER succeeded providing results for the (*Task 1: FMA-NCI small fragments*)`[Precision : 0.96 / Recall : 0.83]` and task 3 of the track (*FMA-SNOMED small fragments*) with a Precision of 0.82 and Recall of 0.42.

2.6 Phenotype

In the Phenotype track, the system succeeded in processing only the DOID-ORDO subcase by identifying 1824 matches for 1237 expected ones, `[Precision : 0.86 / Recall : 0.59]`.

3 Conclusion

In this paper, we briefly described the alignment system KEPLER with comments of the results obtained according to the OAEI 2018 tracks, corresponding to the SEALS platform evaluation modality. Several observations regarding these results were highlighted, in particular the impact of the elimination of any ontological resource on the similarity values. KEPLER is an ongoing work which borrows its idea from two previous systems, CLONA [15] and SERVOMAP [9]. It showed promising results for this second participation. As future work, the idea is to support the instance based ontology alignment in a wider range and contexts [17]. We have dealt with this issue before [18, 19], but the test base update imposes other challenges in terms of the used ontological languages and the evolutive semantic description formalisms.

References

1. Berners-Lee, T.: Designing the web for an open society. In: Proceedings of the 20^{th} International Conference on World Wide Web (WWW2011), Hyderabad, India (2011) 3–4
2. Suchanek, F.M., Varde, A.S., Nayak, R., Senellart, P.: The hidden web, xml and semantic web: A scientific data management perspective. Computing Research Repository (2011) 534–537
3. Kachroudi, M., Ben Moussa, E., Zghal, S., Ben Yahia, S.: Ldoa results for oaei 2011. In: Proceedings of the 6th International Workshop on Ontology Matching (OM-2011) Colocated with the 10^{th} International Semantic Web Conference (ISWC-2011), Bonn, Germany (2011) 148–155

4. Kachroudi, M., Zghal, S., Ben Yahia, S.: Alignement d'ontologies en utilisant des ressources externes : Linked data. In: Actes des 4èmes Journées Francophones sur les Ontologies (JFO'2011), Montréal, Canada (2011) 259–264
5. Diallo, G.: Efficient building of local repository of distributed ontologies. In: Proceedings of the Seventh International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2011, Dijon, France, November 28 - December 1, 2011. (2011) 159–166
6. Dramé, K., Diallo, G., Delva, F., Dartigues, J., Mouillet, E., Salamon, R., Mougin, F.: Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: An application to alzheimer's disease. Journal of Biomedical Informatics **48** (2014) 171–182
7. Euzenat, J., Mocan, A., Scharffe, F.: Ontology alignments: an ontology management perspective. Ontology management: semantic web, semantic web services, and business applications (2008)
8. Kachroudi, M., Zghal, S., Ben Yahia, S.: Ontopart: at the cross-roads of ontology partitioning and scalable ontology alignment systems. International Journal of Metadata, Semantics and Ontologies **8**(3) (2013) 215–225
9. Diallo, G.: An effective method of large scale ontology matching. Journal of Biomedical Semantics **5(44) doi:10.1186/2041-1480-5-44** (2014)
10. Kachroudi, M., Hassen, W., Zghal, S., Ben Yahia, S.: Large ontologies partitioning for alignment techniques scaling. In: Proceedings of the 9^{th} International Conference on Web Information Systems and Technologies (WEBIST), 8-10 May, Aachen, Germany (2013) 165–168
11. Kachroudi, M., Ben Yahia, S., Zghal, S.: Damo - direct alignment for multilingual ontologies. In: Proceedings of the 3^{rd} International Conference on Knowledge Engineering and Ontology Development (KEOD), 26-29 October, Paris, France (2011) 110–117
12. Kachroudi, M., Zghal, S., Ben Yahia, S.: When external linguistic resource supports cross-lingual ontology alignment. In: Proceedings of the 5^{th} International Conference on Web and Information Technologies (ICWIT 2013), 9-12, May, Hammamet, Tunisia (2013) 327–336
13. Kachroudi, M., Zghal, S., Ben Yahia, S.: Using linguistic resource for cross-lingual ontology alignment. International Journal of Recent Contributions from Engineering **1**(1) (2013) 21–27
14. Kachroudi, M., Zghal, S., Ben Yahia, S.: Bridging the multilingualism gap in ontology alignment. International Journal of Metadata, Semantics and Ontologies **9**(3) (2014) 252–262
15. El Abdi, M., Souid, H., Kachroudi, M., Ben Yahia, S.: Clona results for oaei 2015. In: Proceedings of the 12^{th} International Workshop on Ontology Matching (OM-2015) Colocated with the 14^{th} International Semantic Web Conference (ISWC-2015). Volume 1545 of CEUR-WS., Bethlehem (PA US) (2015) 124–129
16. Kachroudi, M., Diallo, G., Ben Yahia, S.: Initiating cross-lingual ontology alignment with information retrieval techniques. In: Actes de la 6ème Edition des Journées sur les Ontologies (JFO'2016), Bordeaux, France (2016) 57–68
17. Kachroudi, M., Diallo, G., Ben Yahia, S.: On the composition of large biomedical ontologies alignment. In: Proceedings of the 7^{th} International Conference on Web Intelligence, Mining and Semantics, WIMS 2017. (2017) 1–10
18. Damak, S., Souid, H., Kachroudi, M., Zghal, S.: Exona results for oaei 2015. In: Proceedings of the 12^{th} International Workshop on Ontology Matching (OM-2015) Colocated with the 14^{th} International Semantic Web Conference (ISWC-2015). Volume 1545 of CEUR-WS., Bethlehem (PA US) (2015) 145–149
19. Zghal, S., Kachroudi, M., Damak, S.: Alignement d'ontologies base d'instances indexées. In: Actes de la 6èmes Edition des Journées Francophones sur les Ontologies (JFO'2016), Bordeaux, France (2016) 69–74