Clona Results for OAEI 2015

Mariem El Abdi, Hazem Souid, Marouen Kachroudi
and Sadok Ben Yahia

Université de Tunis El Manar, Faculté des Sciences de Tunis, LIPAH Programmation Algorithmique et Heuristique, 2092, Tunis, Tunisie;
elabdi.mariam@gmail.com
swdhazem@gmail.com
{marouen.kachroudi,sadok.benyahia}@fst.rnu.tn

Abstract. This paper presents the results of Clona in the Ontology Alignment Evaluation Initiative campaign (OAEI) 2015. We only participated in Multifarm track, since Clona develops specific techniques for aligning multilingual ontologies. We first give an overview of our alignment system; then we detail the techniques used in our contribution to deal with cross-lingual ontology alignment. Last, we present the results with a thorough analysis and discussion, then we conclude by listing some future work on Clona.

1 Presentation of the system

Multilingualism has become an issue of major interest for the Semantic Web community. This process has been accelerated due to a few initiatives which encourage all the active participants to make their data available to the public. Multilingualism is identified as one of the six challenges of the Semantic Web. Consequently, some solutions were proposed at the ontology level, annotation level and the interface level [1].

At the ontology level, the support should be conceived by the ontology designers to create knowledge representations in diverse natural languages. At the annotation level, tools should be developed to assist users in ontologies annotating independently of the natural languages adopted in their design and development. At the interface level, users should be able to have access to the information in natural languages of their own choice, without any linguistic restriction. The absence of the multilingual aspect coverage can be a real handicap during the information exchange in between various services offered by the Semantic Web [2]. So, application fields are more and more numerous and they put in front very specific difficulties. Moreover, the multilingualism coverage allows the reasoning on the context intersections of various ontological representations. In this register, the issue of reasoning on overlapping context domains led to support multilingual information retrieval and digital content management. Multilingual

ontologies alignment is still a little investigated domain in spite of the multiplicity of the alignment methods which remain restricted to monolingual ontologies [3–6].

Clona as a few methods [7–10] meets challenges strictly bound at the linguistic level in the context of multilingual ontology alignment. The driven idea of our new method is to cross the natural language barrier. Clona presents a novel view to improve the alignment accuracy that draws on the information retrieval techniques.

1.1 State, purpose, general statement

The Clona workflow for the OAEI 2015 comprises six different steps, as flagged by Figure 1 : (**i**) Parsing and Pretreatment, (**ii**) Translation, (**iii**) Indexation, (**iv**) Candidate Mappings Identification and (**vi**) Alignment Generation.

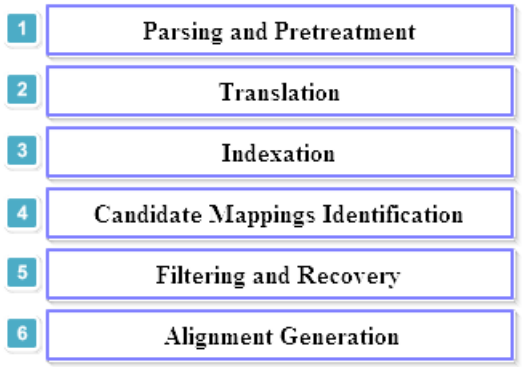

Fig. 1. Clona workflow for OAEI 2015 (Multifarm Track)

Clona is an alignment system which aims through specific techniques to identify the correspondences between two ontologies defined in two different natural languages. Indeed, it starts with a pretreatment stage to model the input ontologies by a format for the rest of the process. The second phase is that of translation into a chosen pivot language and provided by the Microsoft Bing [1] translator.

Thereafter, our method continues with an indexing phase over the considered ontologies. Then these indexes are asked to supply the candidate mappings list to be aligned. Before generating the alignment file, Clona uses a filtering module for recovery and repair.

[1] http://www.microsoft.com/en-us/translator

Parsing and Pretreatment : This phase is crucial for ontologies pretreatment. It is performed using the OWL API. Indeed, it transforms the considered ontologies represented initially as two OWL files in an adequate format for the rest of the treatments. In our case, the goal is to remove all the existing information in both OWL files so that each entity is represented by all its properties. Indeed, the parsing module begins by loading two ontologies to align described in OWL.

This module allows to extract the ontological entities initially represented by a primitive form of lists. In other words, at the parsing stage, we seek primarily to transform an OWL ontology in a well defined structure that preserves and highlight all the information contained in this ontology. Furthermore, in the resulting informative format, has a considerable impact on the results of the similarity computation thereafter. Thus, we get couples formed by the name of the entity and its associated label. In the next step we add an element to such couples to process these entities regardless of their native language.

Translation : The main goal of our approach is to solve the heterogeneity problem mainly due to multilingualism. This challenge brings us to choose between two alternatives, either we consider the translation path to one of the languages according to the two input ontologies, or we consider the translation path to a chosen pivot language. At this stage, we must have a vision of foreseeable rest of our approach. Specifically, at the semantic alignment stage we use an external resource such as WordNet. The latter is a lexical database for the English language. Therefore, our choice is well taken, and we will prepare a translation of the two ontologies to the pivot language, which is English. To perform the translation phase we chose Bing Microsoft tool.

Indexation : Whether on the Internet, with many search engine or local access, we need to find documents or simply sites. Such research is valuable to browse each file and the analysis thereafter. However, the full itinerary of all documents with the terms of a given query is expensive since there are too many documents and prohibitive response times. To enable faster searching, the idea is to execute the analysis in advance and store it in an optimized format for the search. Indexing is one of the novelties of our approach. It consists in reducing the search space through the use of effective search strategy on the built indexes. In fact, we no longer need the sequential scan because with the index structure, we can directly know what document contains a particular word. To ensure this indexing phase we use the Lucene [2] tool. Lucene is a Java API that allows developers to customize and deploy their own indexing and search engine. Lucene uses a suitable technology for all applications that require text search. Indeed, at the end of the indexing process, we get four different indexes to everyone of the two input ontologies depending on the type of the detected entities (*i.e.*, concepts, data types, relationships, and instances). The documents at the indexes represent the se-

[2] https://lucene.apache.org/

mantic information about the entity. These semantic information is obtained by means of an external resource (*i.e.*, WordNet). Indeed, for each entity, Clona keeps the entity name, the label, the label translated to English and its synonyms in English. So with Lucene, we created a set of indexes for the two ontologies, a search query is set up to return all the candidates.

Candidate Mappings Identification : `TermQuery` is the most basic query type to search through an index. It can be built using one term. In our case, `TermQuery`'s role is to find the entities in common between the indexes. Indeed, once the two indexes are set up, the querying step of the latter is activated. Thus, the query implementation satisfies the terminology search and semantic aspects at once as we are querying documents that contain a given ontological entity and its synonyms obtained via WordNet. The result of this process is a set of documents sorted by relevance according to the Lucene score assigned to each returned document. Thus, for each query, Clona keep the first five documents returned and considers them as candidate mappings for the next phase.

Filtering and Recovery : The filtering module consists of two complementary sub modules, each one is responsible of a specific task in order to refine the set of aligned candidates. Indeed, once the list of candidates is ready, Clona uses the first filter. Indeed, we should note that indexes querying may includes a set of redundant mappings. This filter eliminate this redundancy. Indeed, it goes through the list of candidates and for each candidate, it checks if it still exists in the list. If this is the case, it removes the redundant element. At the end of filtering phase, we have a candidates list without redundancy, however, there is always the concern of false positives, indeed, there was the need to establish a second filter. Once the redundant candidates are deleted, Clona uses the second filter that eliminates false positives. This filter is applied to what we call to *partially* redundant entities. An entity is considered *partially* redundant if it belongs to two different mappings (*i.e.*, being given three ontological 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. We note that Clona generates $(1:1)$ alignments. To overcome this challenge, Clona compares the topology of two suspicious entities (e_3 and its neighbor e_4) with respect to the redundant entity e_1 and retains the couple having the highest topological proximity. All candidates following the application of this filter is the subject of alignment file result.

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

1.2 Specific techniques used

Clona has implemented a technique for determining alignment candidates across the power of Lucene search engine. In addition, during the translation phase, we have set up a local translator that is built during the alignment process. This treatment reduces the translation time cost and access to the external resource.

1.3 Link to the system and parameters file

CLONA is an open source ontology matching system and is available through this link (`http://www.mediafire.com/download/f6tacrt82sx316u/CLONA_OAEI_2015.zip`).

2 Results

Our system CLONA has been developed with a unique focus on multilingual ontologies the processing, through Multifarm test base. This 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).

3 General Comments

CLONA obtained an F-measure average of 43% and this, positions it in the second place among methods of the OAEI 2015 campaign. The translation treatment has been successful, especially with the technique of pivot language that reduces all ontological entities to one language, which is English. In addition, the enrichment with WordNet as an external resource, increased produced alignments accuracy. The evaluation was conducted according to two scenarios, as shown in Table 3. The first scenario is significantly better than the second, this is explained by the fact that ontologies share the same structure. Indeed, the structural similarity for ontological entities will be important. These values positioned CLONA in the second place compared to OAEI 2015 participant methods. It should be emphasized that in the case Same Ontologies, and over 45 treated language pairs, CLONA ranked first out of 15 couples. This performance is achieved thanks to the Recall values, which reflect the accuracy of the obtained alignments even in the cross-lingual context [3].

Table 1. F-measure average value for CLONA on Multifarm track for both test scénaios (Same Ontologies and Different Ontologies)

	Same Ontologies	Different Ontologies
	F-measure	**F-measure**
CLONA	**0.58**	**0.39**

[3] More details are available on this link : http://oaei.ontologymatching.org/2015/results/multifarm/index.html

4 Conclusions

Clona participation in OAEI 2015 was encouraging, as it supplies good F-measure values in the two considered scenarios. Results reflects some strengths and some positive aspects.

References

1. Benjamins, V., Contreras, J., Corcho, O., Gómez-Pérez, A.: Six challenges for the semantic web. In: Special Interest Group on Semantic Web and Information Systems (SIGSEMIS Buelletin). (2004)
2. Euzenat, J., Shvaiko, P.: Ontology Matching (Second Edition). Springer-Verlag, Heidelberg (DE) (2013)
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 10th International Semantic Web Conference (ISWC-2011), Bonn, Germany (2011) 148–155
4. Zghal, S., Kachroudi, M., Ben Yahia, S., Mephu Nguifo, E.: OACAS: Ontologies alignment using composition and aggregation of similarities. In: Proceedings of the 1st International Conference on Knowledge Engineering and Ontology Development (KEOD 2009), Madeira, Portugal (2009) 233–238
5. Euzenat, J., Ferrara, A., Meilicke, C., Pane, J., Scharffe, F., Shvaiko, P., Stuckenschmidt, H., Sváb-Zamazal, O., Svátek, V., dos Santos, C.T.: Results of the ontology alignment evaluation initiative 2010. In: Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), Shanghai, China, November 7, 2010. Volume 689 of CEUR-WS. (2010)
6. Euzenat, J., Ferrara, A., van Hage, W.R., Hollink, L., Meilicke, C., Nikolov, A., Ritze, D., Scharffe, F., Shvaiko, P., Stuckenschmidt, H., Sváb-Zamazal, O., dos Santos, C.T.: Results of the ontology alignment evaluation initiative 2011. In: Proceedings of the 6^{th} International Workshop on Ontology Matching (OM-2011), Bonn, Germany, October 24, 2011. Volume 814 of CEUR-WS. (2011)
7. 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
8. Ngo, D., Bellahsene, Z.: Yam++ results for oaei 2012. In: Proceedings of the 9^{th} International Workshop on Ontology Matching (OM-2012) Colocated with the 11^{th} International Semantic Web Conference (ISWC-2012). Volume 946 of CEUR-WS., Boston, USA (2012) 226–233
9. Groß, A., Hartung, M., Kirsten, T., Rahm, E.: Gomma results for oaei 2012. In: Proceedings of the 9^{th} International Workshop on Ontology Matching (OM-2012) Colocated with the 11^{th} International Semantic Web Conference (ISWC-2012). Volume 946 of CEUR-WS., Boston, USA (2012) 133–140
10. Kachroudi, M., Zghal, S., , Ben Yahia, S.: When external linguistic resource supports cross-lingual ontology alignment. In: In Proceedings of the 5^{th} International Conference on Web and Information Technologies (ICWIT 2013), 9-12, May, Hammamet, Tunisia (2013) 327–336