OECM: A Cross-lingual Approach for Ontology Enrichment

Shimaa Ibrahim[1,2], Said Fathalla[1,3], Hamed Shariat Yazdi[1], Jens Lehmann[1,4], and Hajira Jabeen[1]

[1] Smart Data Analytics (SDA), University of Bonn, Germany
ibrahim,fathalla,shariat,jens.lehmann,jabeen@cs.uni-bonn.de
[2] Institute of Graduate Studies and Research, University of Alexandria, Egypt
[3] Faculty of Science, University of Alexandria, Egypt
[4] Enterprise Information Systems Department, Fraunhofer IAIS, Bonn, Germany

Abstract. Due to the rapid expansion of multilingual data on the web, developing ontology enrichment approaches has become an interesting and active subject of research. In this paper, we propose a cross-lingual matching approach for ontology enrichment (OECM) in order to enrich an ontology using another one in a different natural language. A prototype for the proposed approach has been implemented and evaluated using the MultiFarm benchmark. Evaluation results are promising and show higher precision and recall than four state-of-the-art approaches.

Keywords: Cross-lingual matching · Ontology enrichment · Machine translation.

1 Introduction

The increasing amount of multilingual data on the Semantic Web has motivated many researchers to develop ontologies in various natural languages. In fact, ontologies can be enriched by adding additional classes and/or relations extracted from other resources, even in another natural language [7]. Such enrichment is a resource demanding and time-consuming task. Therefore, automated or semi-automated ontology enrichment approaches are highly desired. Most research efforts pay attention to enrich English ontologies, from English resources, rather than non-English ones by applying ontology matching techniques [7]. This raises a key question; How can an ontology be enriched using another ontology in a different natural language? In order to enrich ontologies from multilingual resources, most of the recent efforts in developing different techniques for cross-lingual ontology matching focus on one-to-one translation between ontology concepts [8]. Consequently, inappropriate translations negatively affect the quality of the matching process [8]. Therefore, it is important to develop innovative approaches, which are capable of enriching ontologies by selecting the best translation among all available translations (i.e., one-to-many translations) for a particular term. To the best of our knowledge, only our previous work [1] has addressed the problem of enriching ontologies from the multilingual

text. In this paper, we propose a new approach (OECM) to enrich an ontology, i.e., the target ontology T, using another one, i.e., the source ontology S, in a different natural language. The prominent feature of the proposed approach is the selection of the best translation between all available ones when matching classes among ontologies. This selection has significantly improved the quality of the matching process. Furthermore, the usage of ontologies as the source for the enrichment process can significantly reduce the cost of data pre-processing and cleaning of the data used being used for the enrichment. To evaluate OECM, we compare the cross-lingual ontology matching process with four state-of-the-art approaches. The implementation of OECM and the datasets used for evaluation are publicly available at `https://github.com/shmkhaled/OECM`.

2 The Proposed Approach

Goal: Given two ontologies S and T, in two different natural languages L_1 and L_2 respectively, as RDF triples $\langle s, p, o\rangle \in \mathcal{C} \times \mathcal{R} \times (\mathcal{C} \cup \mathcal{L})$ where $\mathcal{C}$ is the set of ontology domain entities (i.e. classes), $\mathcal{R}$ is the set of relations, and $\mathcal{L}$ is the set of literals. We aim at finding the complementary information $\mathcal{T}_e = S - (S \cap T)$ from S to enrich T.

The methodology of the proposed approach comprises three phases: 1) pre-matching, 2) matching, and 3) enriching. We consider only class labels, or local names, and three standard relations: `rdfs:subClassOf`, `owl:equivalentClass`, `owl:disjointWith`, because there is no need to translate these relations.

1) Pre-matching: T and S are prepared before starting the matching phase by performing two tasks: **a) Pre-processing:** The aim of this task is to prepare the local names and/or labels of classes of S and T by employing several natural language processing techniques, such as tokenization, normalization, stop words removal and POS-tagging. The output of this task is two sets of pre-processed classes $\mathcal{C}'_S$ and $\mathcal{C}'_T$ for S and T respectively, **b) Translation:** Each class in $\mathcal{C}'_S$ is translated using Google Translator to the language of T (i.e., L_2). A list of translations is associated with each class, for example, the class label "Thema" in German, has a list of two English translations: "Subject and Topic". The best translation will be selected in the next phase.

2) Matching: In order to identify which, and where the new information will be added to T, potential matches between S and T should be identified. We use two types of matching: *Terminological matching* and *Structural matching*. **a) Terminological matching:** This task is used to identify which information can be added to T. In order to choose the best translation for each class that matches the corresponding one in T, we perform a pairwise string matching between them. We chose Jaccard similarity as a string similarity metric because it has achieved the best score in terms of precision in the experiments conducted for the ontology alignment task in the MultiFarm benchmark[5] [2]. We consider

[5] `https://www.irit.fr/recherches/MELODI/multifarm/`

similarity scores greater than or equal to a specific threshold θ to get the best matches. After running the experiments for ten times, we obtained the best value of θ which gives the best matching results. If no match is found, this class is considered as a new class, which will be added to T. At the end, matched classes are validated by experts in order to confirm that the best translation is selected for each class. **b) Structural matching:** It is used to identify where the new information can be added to T. Each class in S is replaced by its best translation found in the previous matching in order to get a translated ontology S_{trans}. We apply a pairwise triple comparison between S_{trans} and T to get the set of triples to be enriched $\mathcal{T}_e$, which is represented by $\langle s, p, o, F\rangle$. Each triple is associated with a flag F, with a value either $'E'$ for enrichment or $'A'$ for addition. For a particular triple, if $s \in \mathcal{C}'_T$ and $o \notin \mathcal{C}'_T$, then $F = 'E'$, i.e. this triple is needed to enrich the existing information in T, while if $s \notin \mathcal{C}'_T$ and $o \in \mathcal{C}'_T$, then $F = 'A'$, i.e. this triple is needed to add a new class to T.

3) Enrichment: $\mathcal{T}_e$ is used to enrich T according to the flags associated with each triple. We enrich the Scientific Events Ontology [4] (with 49 classes) (SEO_{en}), which is written in English using the Conference_{de} ontology (with 60 classes) from the MultiFarm dataset (see section 3), which is written in German. OECM has identified new 15 triples to enrich SEO_{en}. For instance, `<ConferenceContributor, subClassOf, Person, 'A'>` is used to add a new class `ConferenceContributor`, as a `subClassOf Person`, to SEO_{en}. In addition, `<KeynoteSpeaker, subClassOf, ConferenceContributor, 'E'>` is used to enrich SEO_{en} with additional information, i.e. adding a new relation `subClassOf` between the two classes. The complete 15 triples can be found at the GitHub repository. We have succeeded to enrich SEO_{en} by 93.75% of the triples identified by an expert.

3 Evaluation

We use ontologies in the MultiFarm benchmark to measure the quality of the cross-lingual matching process. MultiFarm consists of seven ontologies, their translation into nine languages, and the corresponding cross-lingual alignments between them (i.e., the gold standard). We compare our results with four state-of-the-art approaches (see Table 1) for matching Conference_{de} with Ekaw_{en} and Conference_{de} with Edas_{en} ontologies. OECM outperforms all other systems in terms of precision, recall, and F-measure. For AML [3], authors include pre-computed dictionaries with translations, to overcome the query limit of Microsoft Translator which decrease the efficiency of their approach. LogMap [5] depends mainly on the initial mappings to discover new mappings, which decreased after performing the translation. XMap [9] did not achieve satisfactory results because of many internal exceptions. Surprisingly, we found seven new alignments, which did not exist in the gold standard, when matching Conference_{de} with Ekaw_{en}, for instance, (`<LeiterDerWorkshops>`$_{de}$,`<Workshop_Chair>`$_{en}$).

Table 1. State-of-the-art comparison results

Approaches	Conference$_{de} \times Ekaw_{en}$			Conference$_{de} \times Edas_{en}$		
	Precision	Recall	F-measure	Precision	Recall	F-measure
AML [3]	0.56	0.20	0.29	0.86	0.35	0.50
KEPLER [6]	0.33	0.16	0.22	0.43	0.18	0.25
LogMap [5]	0.71	0.20	0.31	0.71	0.29	0.42
XMap [9]	0.18	0.16	0.17	0.23	0.18	0.20
OECM	**0.75**	**0.67**	**0.71**	**0.93**	**0.76**	**0.84**

4 Conclusion

We present a new approach (OECM) in order to enrich ontologies using other ontologies in different natural languages. Terminological and structural matching have been used in order to identify which, and where, information, in the source ontology, can be used to enrich the target ontology. We consider all available translations for each term and select the best one that matches the corresponding term in the target ontology. Such selection has significantly improved the quality of the matching process. It is worth to mention that OECM is able to find new alignments, which were missing in the gold standard. OECM outperforms all other systems in terms of precision, recall, and F-measure. We are in the process of investigating the usage of semantic similarity between terms in the matching process, in addition to considering other non-standard semantic relations and individuals in the enrichment process.

References

1. Ali, M., Fathalla, S., Ibrahim, S., Kholief, M., Hassan, Y.: Cross-lingual ontology enrichment based on multi-agent architecture. Procedia Computer Science **137**, 127–138 (2018)
2. Cheatham, M., Hitzler, P.: String similarity metrics for ontology alignment. In: International Semantic Web Conference. pp. 294–309. Springer (2013)
3. Faria, D., Pesquita, C., Balasubramani, B.S., Tervo, T., Carrio, D., Garrilha, R., Couto, F.M., Cruz, I.F.: Results of aml participation in oaei 2018 p. In Press (2019)
4. Fathalla, S., Vahdati, S., Lange, C., Auer, S.: The scientific events ontology of the openresearch curation platform. In: Proceedings of the Symposium on Applied Computing. pp. 2311–2314. ACM (2019)
5. Jiménez-Ruiz, E., Grau, V.C.: Logmap family participation in the oaei 2018 p. In Press (2019)
6. KACHROUDI, M., DIALLO, G., YAHIA, S.B.: Oaei 2018 results of kepler. In: 13th ISWC workshop on OM. p. In Press (2019)
7. Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., Zavitsanos, E.: Ontology population and enrichment: State of the art. In: Knowledge-driven multimedia information extraction and ontology evolution. pp. 134–166. Springer-Verlag (2011)
8. Trojahn, C., Fu, B., Zamazal, O., Ritze, D.: State-of-the-art in multilingual and cross-lingual ontology matching. In: Towards the Multilingual Semantic Web, pp. 119–135. Springer (2014)

9. Warith Eddine DJEDDI, S.B.Y., KHADIR, M.T.: Xmap results for oaei 2018. In: 13th ISWC workshop on OM. p. In Press (2019)