From Monolingual to Multilingual Ontologies: The Role of Cross-lingual Ontology Enrichment

Conference Paper · July 2019

5 authors, including:

- Shimaa Ibrahim
  University of Bonn
  6 PUBLICATIONS  7 CITATIONS
  SEE PROFILE

- Hamed Shariat Yazdi
  University of Bonn
  20 PUBLICATIONS  38 CITATIONS
  SEE PROFILE

- Said Fathalla
  University of Bonn
  31 PUBLICATIONS  84 CITATIONS
  SEE PROFILE

- Jens Lehmann
  University of Bonn
  273 PUBLICATIONS  10,290 CITATIONS
  SEE PROFILE

Some of the authors of this publication are also working on these related projects:

- LC-QuAD: Large-Scale Complex Question Answering Dataset View project
- DL-Learner View project

All content following this page was uploaded by Said Fathalla on 07 July 2019.
The user has requested enhancement of the downloaded file.
From Monolingual to Multilingual Ontologies:
The Role of Cross-lingual 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. While the multilingual data on the Semantic Web grows rapidly, the building of multilingual ontologies from monolingual ones is still cumbersome and hampered due to the lack of techniques for cross-lingual ontology enrichment. Cross-lingual ontology enrichment greatly facilitates the semantic interoperability between different ontologies in different natural languages. Achieving such enrichment by human labor is very costly and error-prone. Thus, in this paper, we propose a fully automated ontology enrichment approach (OECM), which builds a multilingual ontology by enriching a monolingual ontology from another one in a different natural language, using a cross-lingual matching technique. OECM selects the best translation among all available translations of ontology concepts based on their semantic similarity with the target ontology concepts. We present a use case of our approach for enriching English Scholarly Communication Ontologies using German and Arabic ontologies from the MultiFarm benchmark. We have compared our results with the results from the Ontology Alignment Evaluation Initiative (OAEI 2018). Our approach has higher precision and recall in comparison to five state-of-the-art approaches. Additionally, we recommend some linguistic corrections in the Arabic ontologies in Multifarm which have enhanced our cross-lingual matching results.

Keywords: Cross-lingual ontology enrichment · Cross-lingual matching · multilingual ontology · Ontology engineering · Knowledge management.

1 Introduction

The wide proliferation of multilingual data on the Semantic Web results in many ontologies scattered across the web in various natural languages. According to the Linked Open Vocabularies\(^5\) the majority of the ontologies in the Semantic Web are in English, however, ontologies in other Indo-European languages also exist. For instance, out of a total 666 vocabularies, 482 are in English, 53

\(^5\) https://lov.linkeddata.es/dataset/lov/vocabs
Few ontologies exist in non-Indo-European languages, such as 13 in Japanese and seven in Arabic. Monolingual ontologies with labels or local names presented in a certain language are not easily understandable to speakers of other languages. Therefore, in order to enhance semantic interoperability between monolingual ontologies, approaches for building multilingual ontologies from the existing monolingual ones should be developed [27]. Multilingual ontologies can be built by applying cross-lingual ontology enrichment techniques, which expand the target ontology with additional concepts and semantic relations extracted from external resources in other natural languages [24]. For example, suppose we have two ontologies; Scientific Events Ontology in English (SEO_{en}) and Conference in German (Conference_{de}). Both SEO_{en} and Conference_{de} have complementary information, i.e. SEO_{en} has some information which does not exist in Conference_{de} and vice versa. Let us consider a scenario where a user wants to get information from both SEO_{en} and Conference_{de} to be used in an ontology-based application. This may not be possible without a cross-lingual ontology enrichment solution, which enrich the former by the complementary information in the latter. Manual ontology enrichment is a resource demanding and time-consuming task. Therefore, fully automated cross-lingual ontology enrichment approaches are highly desired [24]. Most of the existing work in ontology enrichment focus on enriching English ontologies from English sources only (monolingual enrichment) [24]. To the best of our knowledge, only our previous work [1,14] has addressed the cross-lingual ontology enrichment problem by proposing a semi-automated approach to enrich ontologies from multilingual text or from other ontologies in different natural languages.

In this paper we address the following research question; how can we automatically build multilingual ontologies from monolingual ones? We propose a fully automated ontology enrichment approach in order to create multilingual ontologies from monolingual ones using cross-lingual matching. We extend our previous work [14] by: 1) using the semantic similarity to select the best translation of class labels, 2) enriching the target ontology by adding new classes in addition to all their related subclasses in the hierarchy, 3) using ontologies in non-Indo-European languages (e.g., Arabic), as the source of information, 4) building multilingual ontologies, and 5) developing a fully automated approach. OECM comprises six phases: 1) translation: translate class labels of the source ontology, 2) pre-processing: process class labels of the target and the translated source ontologies, 3) terminological matching: identify potential matches between class labels of the source and the target ontologies, 4) triple retrieval: retrieve the new information to be added to the target ontology, 5) enrichment: enrich the target ontology with new information extracted from the source ontology, and 6) validation: validate the enriched ontology. A noticeable feature of OECM is that we consider multiple translations for a class label. In addition, the use of semantic similarity has significantly improved the quality of the matching process. We present a use case for enriching the Scientific Events Ontology (SEO) a
scholarly communication ontology for describing scientific events, from German and Arabic ontologies. We compare OECM to five state-of-the-art approaches for cross-lingual ontology matching task. OECM outperformed these approaches in terms of precision, recall, and F-measure. Furthermore, we evaluate the enriched ontology by comparing it against a Gold standard created by ontology experts. The implementation of OECM and the datasets used in the use case are publicly available.

The remainder of this paper is structured as follows: we present an overview of related work in section 2. Overview of the proposed approach is described in section 3. In order to illustrate possible applications of OECM, a use case is presented in section 4. Experiments and evaluation results are presented in section 5. Finally, we conclude with an outline of the future work in section 6.

2 Related Work

A recent review of the literature on multilingual Web of Data found that the potential of the Semantic Web for being multilingual can be accomplished by techniques to build multilingual ontologies from monolingual ones [12]. Multilingual enrichment approaches are used to build multilingual ontologies from different resources in different natural languages [6,25,5]. Espinoza et al. [6] has proposed an approach to generate multilingual ontologies by enriching the existing monolingual ontologies with multilingual information in order to translate these ontologies to a particular language and culture (ontology localization). In fact, ontology enrichment depends on matching the target ontology with external resources, in order to provide the target ontology with additional information extracted from the external resources.

All the literature have focused on the cross-lingual ontology matching techniques which are used for matching different natural languages of linguistic information in ontologies [12,27]. Meilicke et al. [21] created a benchmark dataset (MultiFarm) that results from the manual translations of a set of ontologies from the conference domain into eight natural languages. This dataset is widely used to evaluate the cross-lingual matching approaches [29,16,15,16]. Manual translation of ontologies can be infeasible when dealing with large and complex ontologies. Trojahn et al. [28] proposed a generic approach which relies on translating concepts of source ontologies using machine translation techniques into the language of the target ontology, then they apply monolingual matching approaches to match concepts between the source ontologies and the translated ones. Fu et al. [10,11] proposed an approach to match English and Chinese ontologies by considering the semantics of the target ontology, the mapping intent, the operating domain, the time and resource constraints and user feedback. Hertling and Paulheim [13] proposed an approach which utilizes Wikipedias inter-language links for finding corresponding ontology elements. Lin and Krizhanovsky [18] proposed an approach which use Wiktionary as a source of background knowl-

7 https://github.com/shmkhaled/OECM
8 https://www.wiktionary.org/
edge to match English and French ontologies. Tigrine et al. [26] presented an approach, which relies on the multilingual semantic network BabelNet [9] as a source of background knowledge, to match several ontologies in different natural languages. In the context of OAEI 2018 campaign [10] for evaluating ontology matching technologies, AML [7], KEPLER [16], LogMap [15] and XMap [29] provide high-quality alignments. These systems use terminological and structural alignments in addition to using external lexicon, such as WordNet [11] and UMLS-lexicon [12] in order to get the set of synonyms for the ontology elements. In order to deal with multilingualism, AML and KEPLER rely on getting one translation for a concept (one-to-one translation) using machine translation technologies, such as Microsoft translator, before starting the matching process. LogMap and XMap do not provide any information about the utilized translation methodology. Moreover, LogMap is an iterative process, that starts from initial mappings (almost exact lexical correspondences) to discover new mappings. It is mentioned in [15] that the main weakness of LogMap is that it can not find matching between ontologies which do not provide enough lexical information as it depends mainly on the initial mappings. A good literature of the state-of-the-art approaches in cross-lingual ontology matching is provided in [27].

Most of the literature have focused on enriching monolingual ontologies with multilingual information in order to translate or localize these ontologies. In addition, in the cross-lingual ontology matching task, there is a lack of exact one-to-one translation between terms across different natural languages which negatively affects the matching results. We address this limitations in our proposed approach by building multilingual ontologies, where a class label is presented by several natural languages, from monolingual ones. Such approach support the ontology matching process with multiple translations for a class label in order to enhance the matching results.

3 The Proposed Approach

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

The proposed approach comprises six phases (Figure 1): translation, preprocessing, terminological matching, triple retrieval, enrichment, and validation. The input is the two ontologies in two different natural languages, i.e. the target ontology $T$ and the source ontology $S$. The output is the multilingual enriched ontology $T_{enriched}$ in two different natural languages $L_1$ and $L_2$. In the following subsections, we describe each of these phases in details.

---

[9] https://babelnet.org/
[10] http://oaei.ontologymatching.org/2018/results/multifarm/index.html
[11] https://wordnet.princeton.edu/
[12] https://www.nlm.nih.gov/research/umls/
3.1 Translation

Let $C_S$ and $C_T$ be the set of classes in $S$ and $T$ respectively. Each class is represented by label or a local name. The aim of this phase is to translate each class in $C_S$ to the language of $T$ (i.e. $L_T$). All available translations are considered for each class. Therefore, the output of the translation is $C_{S-translated}$ which has each class, in $S$, associated with a list of all available translations. For example, the class Thema in German has a list of English translations (Subject and Topic), and the class label “مراجعة” in Arabic has a list of English translations such as “Review, Revision, Check”. The best translation will be selected in the terminological matching phase (subsection 3.3).

3.2 Pre-processing

The aim of this phase is to process classes of $C_T$ and lists of translations in $C_{S-translated}$ by employing a variety of natural language processing (NLP) techniques, such as tokenization, POS-tagging (part-of-speech tagging), and lemmatization, to make it ready for the next phases. In order to enhance the similarity results between $C_T$ and $C_{S-translated}$, stop words are removed and normalization methods and regular expressions are used to remove punctuation, symbols, additional white spaces, and to normalize the structure of strings. Furthermore, our pre-processing is capable of recognizing classes such as camel cases “ReviewArticle” and adds a space between lower-case and upper-case letters “Review Article” (i.e. true casing technique [19]). The output of this phase is $C'_T$, which has pre-processed translations of classes in $T$, and $C'_{S-translated}$, which has pre-processed translations for each class in $S$. 

**Fig. 1.** The workflow of OECM.
3.3 Terminological Matching

The aim of this phase is to identify potential matches between class labels of $S$ and $T$. We perform a pairwise lexical and/or semantic similarity between the list of translations of each class in $C_{S-\text{translated}}'$ and $C_T'$ to select the best translation for each class in $S$ that matches the corresponding class in $T$ (see algorithm 1). Jaccard similarity [23] is used to filter the identical concepts instead of using semantic similarity from the beginning because there is no need for extra computations to compute semantic similarity between two identical classes. The reason behind choosing the Jaccard similarity is that according to the experiments conducted for the ontology alignment task for the MultiFarm benchmark in [2], Jaccard similarity has achieved the best score in terms of precision. For non-identical concepts, we compute the semantic similarity using the path length measure, based on WordNet [11], which returns the shortest path between two words in WordNet hierarchy [3]. If two words are semantically equivalent, i.e., belonging to the same WordNet synset, the path distance is 1.00. We use a specific threshold $\theta$ in order to get the set of matched terms (matched classes) $M$. We obtained the best value of $\theta = 0.9$ which has the best matching results after running the experiments for ten times. If no match is found, we consider this class as a new class that can be added to $T$ and we consider its list of translations as synonyms for that class. Generally, class labels have more than one word, for example “InvitedSpeaker”, therefore, the semantic similarity between sentences presented in [22] is adapted as described in algorithm 1 - line 9. Given two sentences $\text{sentence1}$ and $\text{sentence2}$, the semantic similarity of each sentence with respect to the other is defined by: for each word $w_i \in \text{sentence1}$, the word $w_j$ in $\text{sentence2}$ that has the highest path similarity with $w_i$ is determined. The word similarities are then summed up and normalized with the number of similar words between the two sentences. Next, the same procedure is applied to start with words in $\text{sentence2}$ to identify the semantic similarity of $\text{sentence2}$ with respect to $\text{sentence1}$. Finally, the resulting similarity scores

![Diagram](image_url)
Algorithm 1: Terminological Matching

Data: $C'_{S-translated}$, $C'_T$, $\theta$ similarity threshold
Result: $M$ matched terms, $C'_{S-translated}$

1. foreach $c_s \in C'_{S-translated}$, $t \in listOfTranslations$, $c_t \in C'_T$ do
2.  similarityScore $\leftarrow$ getSimilarity($t, c_t$)
3.  if similarityScore $\geq \theta$ then
4.     $M ::= (t, c_t)$
5.     $C'_{S-translated} = update(C'_{S-translated}, M)$

6. Function getSimilarity(sentence1, sentence2): double
7.     similarity $\leftarrow$ getJaccardSimilarity(sentence1, sentence2)
8.     if similarity $\neq 1$ then
9.         similarity $\leftarrow$ (sentenceSimilarity(sentence1, sentence2) + sentenceSimilarity(sentence2, sentence1))/2
10.    return similarity

11. Function sentenceSimilarity(sentence1, sentence2): double
12.     simScore $\leftarrow$ 0.0
13.     count $\leftarrow$ 0.0
14.     foreach $w_i \in$ sentence1.split(" ") do
15.         foreach $w_j \in$ sentence2.split(" ") do
16.             pathSim $::=$ getPathSimilarity($w_i$, $w_j$)
17.             simScore $+$ = pathSim.max
18.         count $+$ = 1
19.     simScore $\leftarrow$ simScore/count
20.    return simScore

The output of this phase is the list of matched terms $M$ between $C'_T$ and the updated $C'_{S-translated}$.

3.4 Triple Retrieval

The aim of this phase is to identify which and where the new information can be added to $T$. Each class in $S$ is replaced by its best translation found in $C'_{S-translated}$ from the previous phase in order to get a translated ontology $S'_{translated}$ (see Algorithm 2). We design an iterative process in order to obtain $T_c$, which is represented by $(s, p, o)$, that has all possible multilingual information from $S$ to be added to $T$. We initiate the iterative process with all matched terms ($newClasses = M$) in order to get all related classes, if exist. The iterative process has three steps: 1) for each class $c \in newClasses$, all triples $tempTriples$ are retrieved from $S'_{translated}$ where $c$ is a subject or an object, 2)
Algorithm 2: Triple Retrieval

Data: $S, C'_S - \text{translated}, C'_T, M$
Result: $T_e$ triples to be enriched
1 $S_{\text{translated}} \leftarrow \text{translateOntologyClasses}(S, C'_S - \text{translated})$
2 $\text{newClasses} \leftarrow M$
3 while $\text{newClasses.isEmpty()}$ do
4 $\text{tempTriples} \leftarrow \text{getTriplesForNewClasses}(S_{\text{translated}}, \text{newClasses})$
5 $\text{newClasses} \leftarrow \text{getClasses}(\text{tempTriples}).\text{subtract(newClasses)}$
6 $\text{newTriples} \leftarrow \text{newTriples}.\text{union(tempTriples)}$
7 $\text{otherLangTriples} \leftarrow \text{getOtherLangTriples(\text{newTriples}, C'_S - \text{translated})}$
8 $T_e \leftarrow \text{newTriples}.\text{union(\text{foreignLanguageTriples})}$

a new list of new classes is obtained from $\text{tempTriples}$, 3) $\text{tempTriples}$ is added to $\text{newTriples}$ which will be added to $T$. These three steps are repeated until no new classes can be found ($\text{newClasses.isEmpty()} = \text{true}$). Next, we retrieve all available information from the other language for each class in $\text{newTriples}$ such as $\langle \text{president, rdfs:label, "رئيس"}\rangle@\text{ar}$. The output of this phase is $T_e$ which contains all multilingual triples (i.e., in $L_s$ and $L_t$ languages) to be added to $T$.

3.5 Enrichment

The aim of this phase is to enrich $T$ using triples in $T_e$. By using OECM, the target ontology can be enriched from several ontologies in different natural languages sequentially, i.e. one-to-many enrichment. In this case, $T_{\text{enriched}}$ can have more than two natural languages. For instance, English $T$ can be enriched from a German ontology, then the enriched ontology can be enriched again form a different Arabic ontology, i.e. the final result for $T_{\text{enriched}}$ is presented in English, German, and Arabic. With the completion of this phase, we have successfully enriched $T$ and create a multilingual ontology from monolingual ones.

3.6 Validation

The aim of this phase is to validate the enriched ontology, which is a crucial step to detect inconsistencies and syntax errors, which might be produced during the enrichment process [8]. There are two types of validations: syntactic and semantic validation. In the syntactic validation, we validate $T_{\text{enriched}}$ to conform with the W3C RDF standards using the online RDF validation service\(^{13}\) which detects syntax errors, such as missing tags. In the semantic validation, we use two types of reasoners: FaCT++\(^{14}\) and HermiT\(^{15}\) in order to detect inconsistencies in $T_{\text{enriched}}$.

\(^{13}\)https://www.w3.org/RDF/Validator/
\(^{14}\)https://github.com/ethz-asl/libfactplusplus
\(^{15}\)https://github.com/phillord/hermit-reasoner
4 Use Case: Enriching the Scientific Events Ontology

In this use case, we use an example scenario to enrich the SEO\textsubscript{en} ontology (with 49 classes), in English, using the MultiFarm dataset (see section 5). We use the Conference ontology (60 classes) and the ConfOf ontology (38 classes), in German and Arabic respectively, as source ontologies. This use case aims to show the whole process starting from submitting the source and target ontologies until producing the enriched multilingual ontology. Here, the source ontology is the German ontology Conference\textsubscript{de} and the target ontology is the English ontology SEO\textsubscript{en}. The output is the enriched ontology SEO\textsubscript{en−de}, which becomes a multilingual ontology in English and German. Table 1 demonstrates the enrichment process for SEO\textsubscript{en} from Conference\textsubscript{de} and shows the output sample of each phase starting from the translation phase to the produced set of triples which are used to enrich SEO\textsubscript{en}. In the terminological matching task, the relevant matching results (with similarity scores in bold) are identified with $\theta \geq 0.9$. The iterative process, in the triple retrieval phase, is initiated with the identified matched terms, for example, person class. At the first iteration, six triples (not all results are exist in the table because of the limited space) are produced such as \((\text{conference contributor}, \text{rdfs:subClassOf}, \text{person})\), where the matched term person is located at the object position. New classes are determined from the produced triples such as conference contributor and committee member (in bold). At the second iteration, all triples that have these new classes, as subject or object, are retrieved, for example: for the committee member class, the triples \((\text{committee member}, \text{rdf:type}, \text{Class})\) and \((\text{chairman}, \text{rdfs:subClassOf}, \text{committee member})\) are retrieved. This process is repeated again and new classes are identified from the produced triples such as chairman.
Table 1. Use case: the sample output of each phase, from translation to triple retrieval.

| Phase                        | Output                                                                 |
|------------------------------|------------------------------------------------------------------------|
| Translation                  | (Thema)\textsubscript{de} \rightarrow (subject, topic)\textsubscript{en} |
|                              | (Gutachter)\textsubscript{de} \rightarrow (reviewer, expert)\textsubscript{en} |
|                              | (Herausgeber)\textsubscript{de} \rightarrow (publisher, editor)\textsubscript{en} |
|                              | (Fortschritte der Konferenz)\textsubscript{de} \rightarrow (Progress of the conference)\textsubscript{en} |
| Pre-processing               | SizeOrDuration \rightarrow size duration                              |
|                              | WorkshopProposals \rightarrow workshop proposal                        |
|                              | InvitedSpeaker \rightarrow invite speaker                              |
|                              | In-useTrack \rightarrow use track                                       |
| Terminological matching score results | (invited speaker, keynote speaker, 0.57)                             |
|                              | (person, person, 1.00)                                                 |
|                              | (tutorial, tutorial proposals, 0.78)                                   |
|                              | (prize, award, 1.00)                                                   |
|                              | (conference document, license document, 0.61)                          |
|                              | (publisher, publisher, 1.00)                                           |
|                              | (conference series, event series, 0.79)                                |
|                              | (conference series, symposium series, 0.75)                            |
|                              | (proceedings, proceedings, 1.00)                                       |
|                              | (poster, posters track, 0.78)                                          |
| Triple Retrieval (Iterative process) | \(T_e\) \rightarrow conference contributor, rdfs:subClassOf, person |
| 1\textsuperscript{st} Iteration: | \(T_e\) \rightarrow committee member, rdfs:subClassOf, person         |
| 2\textsuperscript{nd} Iteration: | \(T_e\) \rightarrow committee member, rdf:type, Class                |
|                              | \(T_e\) \rightarrow chairman, rdfs:subClassOf, committee member       |
|                              | \(T_e\) \rightarrow invited speaker, rdfs:subClassOf, conference contributor |
|                              | \(T_e\) \rightarrow regular author, rdfs:subClassOf, conference contributor |
| Triple Retrieval (\(T_e\))   | \(T_e\) \rightarrow committee member, rdf:type, Class                |
|                              | \(T_e\) \rightarrow committee member, rdfs:label, "committee member"\textsubscript{en} |
|                              | \(T_e\) \rightarrow committee member, rdfs:label, "Angehörige des Ausschusses"\textsubscript{de} |
|                              | \(T_e\) \rightarrow chairman, rdfs:subClassOf, committee member       |

The iterative process ended at the fifth iteration where three triples are produced without any new classes. The output of this phase is \(T_e\) which has 40 new triples (with 20 new classes and their German labels), to be added to SEO\textsubscript{en} and produce SEO\textsubscript{en--de}. Figure 3 shows a small fragment of the enriched ontology SEO\textsubscript{en--de}, in Turtle, after completing the enrichment process. The resulting multilingual ontology contains a newly added class CommitteeMember with its English and German labels, a new relation rdfs:subClassOf between the two classes CommitteeMember and Chair, and new German labels such as Herausgeber and Vorsitzender for classes Publisher and Chair respectively. Similarly, SEO\textsubscript{en--de} is enriched from the Arabic ontology ConfOf\textsubscript{ar}, where all classes with English labels in SEO\textsubscript{en--de} are matched with class labels in ConfOf\textsubscript{ar}. The produced SEO\textsubscript{en--de--ar} has 113 new triples with 37 new classes.
with their Arabic labels. Final output results can be found at the GitHub repository.

5 Evaluation

The aim of this evaluation is to measure the quality of the cross-lingual matching process in addition to the enrichment process. We use ontologies in MultiFarm benchmark, a benchmark designed for evaluating cross-lingual ontology matching systems. MultiFarm consists of seven ontologies (Cmt, Conference, ConfOf, Edas, Ekaw, Iasted, Sigkdd) originally coming from the Conference benchmark of OAEI, their translation into nine languages (Chinese, Czech, Dutch, French, German, Portuguese, Russian, Spanish and Arabic), and the corresponding cross-lingual alignments between them.

5.1 Experimental Setup

All phases of OECM have been implemented using Scala and Apache Spark. SANSA-RDF library with Apache Jena framework are used to parse and manipulate the input ontologies (as RDF triples). Google Translator is used to translate class labels/local names of source ontologies. In order to process the class labels, the Stanford CoreNLP is used. All experiments are carried out on Ubuntu 16.04 LTS operating system with an Intel Corei7-4600U CPU @ 2.10GHz x 4 CPU and 10 GB of memory. In our experiments, we consider English ontologies as target ontologies to be enriched from German and Arabic ontologies.

Our evaluation has three tasks: 1) evaluating the effectiveness of the cross-lingual matching process in OECM compared to the reference alignment provided in the MultiFarm benchmark, 2) comparing OECM matching results with four state-of-the-art approaches, in addition to our previous work (OECM 1.0), and 3) evaluating the quality of the enrichment process.

5.2 Effectiveness of OECM

In this experiment, we use the English version of Cmt ontology as the source ontology, and German and Arabic versions of Conference, ConfOf, and Sigkdd ontologies as target ontologies. We match class labels in Cmt ontology with class labels of German and Arabic versions of Conference, ConfOf, and Sigkdd ontologies separately. The resulting alignments are compared with the reference alignments, as a gold standard, provided in the benchmark for each pair of
Table 2. Precision, recall and F-measures for the cross-lingual matching

| Ontology pairs     | German × English | Arabic × English |
|--------------------|------------------|------------------|
|                    | Precision | Recall | F-measure | Precision | Recall | F-measure | Precision | Recall | F-measure |
| Conference × Cmt   | 1.00      | 0.38   | 0.56      | 1.00      | 0.33   | 0.42      | 0.50      | 0.59   |
| ConfOf × Cmt       | 1.00      | 0.70   | 0.82      | 1.00      | 0.30   | 0.60      | 0.46      | 0.75   |
| Sigkdd × Cmt       | 1.00      | 0.90   | 0.95      | 1.00      | 0.40   | 0.80      | 0.57      | 0.89   |

ontologies. Table 2 shows the precision, recall and F-measure of the matching process for each pair of ontologies. OECM achieves the highest precision of 1.00 for all pair of ontologies. Meanwhile, OECM achieves the highest recall and F-measure of 0.90 and 0.95 respectively for matching the German Sigkdd with the English Cmt. As two authors of this work are native speakers of Arabic, we found some linguistic mistakes in the Arabic ontologies which negatively affect the translation and the matching results. Therefore, we correct these mistakes and make it available at the GitHub repository. Matching results before and after the corrections are presented in the table, where such corrections have greatly improved the matching results in terms of recall and F-measure. For instance, in matching the Arabic Sigkdd with the English Cmt, recall and F-measure are enhanced by 40% and 32% respectively.

5.3 Comparison with the state-of-the-art

We identified four of the related approaches (AML, KEPLER, LogMap, and XMap) to be included in our evaluation in addition to OECM 1.0. The other related work, neither publish their code, nor their evaluation datasets [26,11,10]. In order to compare our results with the state-of-the-art, we use German (Conference\textsubscript{de}) and Arabic (Conference\textsubscript{ar}) versions of the Conference ontology as the source ontologies, and Ekaw\textsubscript{en} and Edas\textsubscript{en} ontologies as the target English ontologies. We choose Ekaw\textsubscript{en} and Edas\textsubscript{en} ontologies in this evaluation because they are used in the state-of-the-art systems for evaluation, as mentioned in the results of OAEI 2018\textsuperscript{10}. We generate the gold standard alignments between each pair of ontologies using the Alignment API\textsuperscript{22} in order to compute precision, recall, and F-measures. Table 3 shows the comparison between our results against four state-of-the-art approaches and OECM 1.0 (results for matching English and German ontologies only). In addition, we add the updated Arabic ontology (Conference\textsuperscript{ar}) with our linguistic correction in the matching process in order to show the effectiveness of such corrections. The current version of OECM (OECM 1.1) outperforms all other systems in precision, recall and F-measure. For instance, when matching Conference\textsubscript{de} × Ekaw\textsubscript{en}, OECM 1.1 outperforms LogMap, the highest precision, recall and F-measure among the others, by 29%, 60% and 58% in terms of precision, recall and F-measure respectively. The use of semantic similarity in OECM 1.1 significantly improves
Table 3. State-of-the-art comparison results. Bold entries are the top scores.

| Approaches | Conference\textsubscript{de} × Ekaw\textsubscript{en} | Conference\textsubscript{de} × Edas\textsubscript{en} | Conference\textsubscript{ar} × Ekaw\textsubscript{en} | Conference\textsubscript{ar} × Edas\textsubscript{en} | Conference\textsuperscript{'}\textsubscript{ar} × Ekaw\textsubscript{en} | Conference\textsuperscript{'}\textsubscript{ar} × Edas\textsubscript{en} |
|------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|            | Precision | Recall  | F-measure | Precision | Recall  | F-measure | Precision | Recall  | F-measure | Precision | Recall  | F-measure |
| AML [7]    | 0.56     | 0.20    | 0.29     | 0.86     | 0.35    | 0.50     | 0.64     | 0.39    | 0.28     | 0.71     | 0.42    | 0.29     |
| KEPLER [16]| 0.33     | 0.16    | 0.22     | 0.43     | 0.18    | 0.25     | 0.40     | 0.30    | 0.24     | 0.40     | 0.30    | 0.24     |
| LogMap [15]| 0.71     | 0.20    | 0.31     | 0.71     | 0.29    | 0.42     | 0.40     | 0.13    | 0.08     | 0.40     | 0.18    | 0.12     |
| XMap [29]  | 0.18     | 0.16    | 0.17     | 0.23     | 0.18    | 0.20     | 1.00     | 0.00    | 0.00     | 1.00     | 0.00    | 0.00     |
| OECM 1.0 [14]| 0.75    | 0.67    | 0.71     | 0.93     | 0.76    | 0.84     | 1.00     | 0.80    | 0.89     | 1.00     | 0.78    | 0.88     |
| OECM 1.1   | **1.00** | **0.80**| **0.89** | **1.00** | **0.78**| **0.88** | **1.00** | **0.70**| **0.78** | **1.00** | **0.78**| **0.88**|

the matching results compared to the results of OECM 1.0. For instance, when matching Conference\textsubscript{de} × Ekaw\textsubscript{en}, matching results in OECM 1.0 have been enhanced by 25%, 13%, and 18% in terms of precision, recall and F-measure respectively. When matching Conference\textsubscript{ar} × Edas\textsubscript{en}, XMap outperform OECM by 14% in terms of precision, while OECM outperforms it in both recall and f-measure. It is observed that the precision of OECM slightly decreased because of the linguistic mistakes found in Conference\textsubscript{ar}. When considering Conference\textsuperscript{'}\textsubscript{ar}, which has the linguistic correction, as a source ontology in this matching, the matching results are improved.

5.4 Evaluating the Enrichment Process

According to [4], the enriched ontology can be evaluated by comparing it against a predefined reference ontology (Gold standard). In this experiment, we evaluate the enriched ontology SEO\textsubscript{en–de} (from section 4). A gold standard is created, by ontology experts, by enriching the target ontology manually using Conference\textsubscript{de}. By comparing SEO\textsubscript{en–de} with the gold standard, OECM achieves 1.00, 0.80, and 0.89 in terms of precision, recall, and F-measure respectively. This finding confirms the usefulness of our approach in cross-lingual ontology enrichment.

6 Conclusion

We present a fully automated approach, OECM, for building multilingual ontologies. The strength of our contribution lies in building such ontologies from monolingual ones using cross-lingual matching between ontologies concepts. Indo and non-Indo-European languages resources are used for enrichment in order to illustrate the robustness of our approach. Considering multiple translations of
concepts and the use of semantic similarity measures for selecting the best translation have significantly improved the quality of the matching process. Iterative triple retrieval process has been developed to determine which information, from the source ontology, can be added to the target ontology, and where such information should be added. We show the applicability of OECM by presenting a use case for enriching an ontology in the scholarly communication domain. The results of the cross-lingual matching process are found promising compared to five state-of-the-art approaches, involving the previous version of OECM. Furthermore, evaluating the quality of the enrichment process emphasizes the validity of our approach. Finally, we propose some linguistic corrections for the Arabic ontologies in the MultiFarm benchmark that used in our experiment, which considerably enhanced the matching results. In conclusion, our approach provides a springboard for a new way to build multilingual ontologies from monolingual ones.

In the future, we intend to further consider properties and individuals in the enrichment process. In addition, we aim to apply optimization methods in order to evaluate the efficiency of OECM when enriching very large ontologies.

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. Cross, V.: Semantic similarity: A key to ontology alignment. In: Ontology Matching: OM-2018: Proceedings of the ISWC Workshop. p. 61 (2018)
4. Dellschaft, K., Staab, S.: On how to perform a gold standard based evaluation of ontology learning. In: International Semantic Web Conference. vol. 4273, pp. 228–241. Springer (2006)
5. Embley, D.W., Liddle, S.W., Lonsdale, D.W., Tijerino, Y.: Multilingual ontologies for cross-language information extraction and semantic search. In: International Conference on Conceptual Modeling. pp. 147–160. Springer (2011)
6. Espinoza, M., Gómez-Pérez, A., Mena, E.: Enriching an ontology with multilingual information. In: European Semantic Web Conference. pp. 333–347. Springer (2008)
7. 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)
8. Fathalla, S., Lange, C.: Eventskg: a knowledge graph representation for top-prestigious computer science events metadata. In: International Conference on Computational Collective Intelligence. pp. 53–63. Springer (2018)
9. Fathalla, S., Vahtadi, S., Auer, S., Lange, C.: The scientific events ontology of the openresearch.org curation platform. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. pp. 2311–2313. SAC '19, ACM (2019)
10. Fu, B., Brennan, R.: Cross-lingual ontology mapping and its use on the multilingual semantic web. MSW 571, 13–20 (2010)
11. Fu, B., Brennan, R., OSullivan, D.: A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes. Web Semantics: Science, Services and Agents on the World Wide Web 15, 15–36 (2012)
12. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data. Web Semantics: Science, Services and Agents on the World Wide Web 11, 63–71 (2012)
13. Hertling, S., Paulheim, H.: Wikimatch: using wikipedia for ontology matching. Ontology Matching 946 (2012)
14. Ibrahim, S., Fathalla, S., Yazdi, H.S., Lehmann, J., Jabeen, H.: Oecm: A cross-lingual approach for ontology enrichment. In: European Semantic Web Conference. p. In Press. Springer (2019)
15. Jiménez-Ruiz, E., Grau, V.C.: Logmap family participation in the oaei 2018 p. In Press (2019)
16. KACHROUDI, M., DIALLO, G., YAHIA, S.B.: Oaei 2018 results of kepler. In: 13th ISWC workshop on ontology matching. p. In Press (2019)
17. Lehmann, J., Sejdiu, G., Bühmann, L., Westphal, P., Stadler, C., Ermilov, I., Bin, S., Chakraborty, N., Saleem, M., Ngomo, A.C.N., et al.: Distributed semantic analytics using the sansa stack. In: International Semantic Web Conference. pp. 147–155. Springer (2017)
18. Lin, F., Krizhanovsky, A.: Multilingual ontology matching based on wiktionary data accessible via sparql endpoint. In: RCDL (2011)
19. Lita, L.V., Ittycheriah, A., Roukos, S., Kambhatla, N.: Truecasing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, pp. 152–159. Association for Computational Linguistics (2003)
20. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014), http://www.aclweb.org/anthology/P/P14/P14-5010
21. Melicke, C., García-Castro, R., Freitas, F., Van Hage, W.R., Montiel-Ponsoda, E., De Azevedo, R.R., Stuckenschmidt, H., ŠváB-Zamazal, O., Svátek, V., Tamilin, A., et al.: Multifarm: A benchmark for multilingual ontology matching. Web Semantics: Science, Services and Agents on the World Wide Web 15, 62–68 (2012)
22. Mihalcea, R., Corley, C., Strapparava, C., et al.: Corpus-based and knowledge-based measures of text semantic similarity. In: AAAI. vol. 6, pp. 775–780 (2006)
23. Niwattanakul, S., Singthongchai, J., Naemudorn, E., Wanapu, S.: Using of jaccard coefficient for keywords similarity. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. vol. 1 (2013)
24. 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)
25. Spohr, D., Hollink, L., Cimiano, P.: A machine learning approach to multilingual and cross-lingual ontology matching. In: International Semantic Web Conference. pp. 665–680. Springer (2011)
26. Tigrine, A.N., Bellahsen, Z., Todorov, K.: Light-weight cross-lingual ontology matching with lynam++. In: OTM Confederated International Conferences” On the Move to Meaningful Internet Systems”. pp. 527–544. Springer (2015)
27. 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)
28. Trojahn, C., Quaresma, P., Vieira, R.: A framework for multilingual ontology mapping (2008)
29. Warith Eddine DJEDDI, S.B.Y., KHADIR, M.T.: Xmap results for oaei 2018. In: 13th ISWC workshop on ontology matching. p. In Press (2019)