An API for Multi-lingual Ontology Matching

Cássia Trojahn, Paulo Quaresma, Renata Vieira

INRIA & LIG, Universidade de Évora, PUC-RS
Avenue de l'Europe 655 Montbonnot Saint Ismier - France
Rua Romão Ramalho 59 Évora - Portugal
Av. Ipiranga 6681 Porto Alegre - Brazil
cassia.trojahn@inrialpes.fr, pq@di.uevora.pt, renata.vieira@pucrs.br

Abstract

Ontology matching consists of generating a set of correspondences between the entities of two ontologies. This process is seen as a solution to data heterogeneity in ontology-based applications, enabling the interoperability between them. However, existing matching systems are designed by assuming that the entities of both source and target ontologies are written in the same languages (English, for instance). Multi-lingual ontology matching is an open research issue. This paper describes an API for multi-lingual matching that implements two strategies, *direct translation-based* and *indirect*. The first strategy considers direct matching between two ontologies (i.e., without intermediary ontologies), with the help of external resources, i.e., translations. The *indirect alignment* strategy, proposed by (Jung et al., 2009), is based on composition of alignments. We evaluate these strategies using simple string similarity based matchers and three ontologies written in English, French, and Portuguese, an extension of the OAEI benchmark test 206.

1. Introduction

Ontology Matching is seen as the solution to data heterogeneity in ontology-based applications. Matching ontologies consists of finding corresponding entities (i.e., classes, properties, or instances) in different ontologies (usually one source ontology and one target ontology).

Different systems and algorithms implementing this process have been proposed, which are surveyed from different perspectives in (Euzenat and Shvaiko, 2007). The distinction between them is accentuated by the manner in which they exploit the features within an ontology. Whereas syntactic techniques consider measures of string similarity; semantic ones consider semantic relations usually on the basis of lexical oriented linguistic resources; and structural techniques consider term positions in the ontology hierarchy.

Most ontology matching systems are designed by assuming that the entities of both source and target ontologies are written in the same language (English, for instance)[1]. With the increasing number of distributed resources, services, and applications on the web, multi-lingual ontology matching is likely to become essential.

Few works exploiting this problem have been proposed (see (Jung et al., 2009) and (Fu et al., 2009)). It is corroborated by the number of systems participating in the multi-lingual directories track[2] of Ontology Alignment Evaluation Initiative (OAEI)[3] campaigns. Only one system had participated in 2008 in the test case Japonese-Portuguese alignment and in 2009 there were no participants.

Few works exploiting this problem have been proposed (see (Jung et al., 2009), (Fu et al., 2009) and (Wang et al., 2009)) and there are no public available resources (APIs, tools, and test cases) to be reused. This paper presents a minimal API for multi-lingual matching that implements two strategies, *direct translation-based* and *indirect*. The first strategy considers direct matching between two ontologies (i.e., without intermediary ontologies), with the help of external resources, i.e., translations. The *indirect alignment* strategy, proposed by (Jung et al., 2009), is based on composition of alignments. We provide an implementation for both approaches. Regarding new test cases, one Portuguese ontology based on the OAEI benchmark test 206 is created together with its reference alignments to English and French ontologies. We make such resources publicly available (API, implementation and test cases) in order to contribute for enriching the multi-lingual matching resources.

The main contribution of this paper is to provide and evaluate practical and reusable alternatives for multi-lingual matching, where existing systems are designed to work on specific languages and resouces (e.g. bilingual dictionaries) or depend on intermediary alignments, which are not always available.

The rest of the paper is structured as follows. Section 2. introduces the matching process in general. In Section 3. the strategies for multi-lingual ontology matching are presented. Section 4. details the API and the corresponding implementation. Section 5. describes the Portuguese ontology we create to evaluate the strategies and presents the preliminary evaluation results. Finally, Section 6. concludes the paper and presents the future work.

2. Matching Process

The ontology matching process consists of generating an alignment (A') from a pair of ontologies (o_s and o_t, source and target, respectively). This general definition can be extended by considering additional parameters, such as an input alignment (A) which is to be completed by the process, the alignment parameters (which can be weights, for instance) and some external resources used by the alignment process (e.g., lexicons and databases). This process can be

[1] The ontologies are supposed to have no specific language. For example, an ontology class can be described by different labels that can be written in different languages. In this paper we assume that such descriptions are written in the same language.

[2] http://oaei.ontologymatching.org/2008/mldirectory/

[3] http://oaei.ontologymatching.org/

defined as follows (Euzenat and Shvaiko, 2007):

Definition 1 (Matching process) *The matching process can be seen as a function f which, from a pair of ontologies o_s and o_t to align, an input alignment A, a set of parameters p, and a set of oracles and resources r, returns a new alignment A' between these ontologies:*

$$A' = f(o_s, o_t, A, p, r)$$

An alignment A' is a set of correspondences:

Definition 2 (Correspondence) *Given two ontologies, o_s and o_t, a correspondence is a quadruple:*

$$< e_s, e_t, r, c >$$

where $e_s \in o_s$, $e_t \in o_t$, r is the relation between e_s and e_t, $r \in R$, a set of alignment relations (i.e., $\equiv$, $\sqsubseteq$, or $\sqsupseteq$), and $c \in [0,1]$ is a confidence level (i.e., measure of confidence in the fact that the correspondence holds).

3. Generating Multi-lingual Alignment

The process of generating alignments involves to use techniques to align two ontologies and strategies to combine and treat these alignments. In this paper we discuss on strategies for matching multi-lingual ontologies as a way for exploiting additional steps and resources that can help in this process. This section describes the two strategies we have presented in Section 1..

For the direct translation-based strategy, the entities (labels) of the source ontology are translated into the language of the target ontology, what generates a translated ontology. Then different matchers compute the alignment between the translated and target ontologies (e.g. labels of entities are matched using an edit distance measure). Following the composition indirect based strategy, alignments previously computed by matchers are used to derive new alignments. These strategies are detailed in the following.

3.1. Direct translation-based alignment

The notion of direct translation-based alignment proposed in this paper is a simplification of our previous work (Trojahn et al., 2008), which uses external resources such as WordNet and dictionaries. First, a bilingual dictionary is used to translate each label into the target language. Next, WordNet is used to obtain the set of synonyms for each translated label. For instance, the Portuguese label "Tese" (Thesis) is translated into English using the dictionary and the translation "Thesis" is used to retrieve the corresponding synonymous in WordNet. So, "Tese" is composed by the disjunction of its synonymous in the target language, "Tese" $\equiv$ (Thesis $\sqcup$ Dissertation $\sqcup$ ... $\sqcup$ $t_{syn,n}$). In this paper we consider a direct translation of labels based only on equivalence relations (i.e., Tese $\equiv$ Thesis), using available resources to provide the translations. We do not use WordNet to retrieval the synonymous of each label.

We define the direct translation-based strategy as follows:

Definition 3 (Direct translation-based Alignment) *Given two ontologies, o_s and o_t, written in the languages L_s and L_t, respectively, a direct translation-based alignment A is a set of correspondences:*

$$A_{s,t} = \{< e_i, e_t, r, c >\}$$

where e_i = translate(e_s,L_s,L_t), $e_s \in o_s$, and $e_t \in o_t$.

3.2. Composition-based Indirect Alignment

The indirect alignment by composition is proposed by (Jung et al., 2009). The basic idea is to use intermediary alignments between source and target ontologies and compose one new alignment using such objects. Following this approach, an alignment between French and Portuguese ontologies can be composed by using intermediary alignments in English, i.e., French – English and English – Portuguese alignments.

Definition 4 (Composition (Jung et al., 2009)) *Given two alignments $A_{s,i}$ and $A_{i,t}$, if there exist a certain bridging entity connecting two multi-lingual correspondences, the composed alignment $A_{s,t}$ is given by a set of composed correspondences:*

$$\begin{aligned} A_{(s,t)} &= A_{s,i}.A_{i,t} \\ &= \{< e, e''', F_{rel}(r, r'), F_{conf}(n, n') >\} \end{aligned}$$

where

- $e \in o_s$, $e' \in o_i$, $e'' \in o_i$, $e''' \in o_t$
- $<e,e',r,n> \in A_{s,i}$, $<e'',e''',r',n'> \in A_{i,t}$
- *the bridging entity $e' \equiv e''$ (or $e' \sqsubseteq e''$, $e' \sqsupseteq e''$) and*
- *F_{rel} and F_{conf} are functions for composing two relations and two confidence values, respectively.*

Regarding relation composition, F_{rel}, a composition table stating relation algebra for determining the composed relations between the given two relations of correspondences can be specified as proposed by (Euzenat, 2008). For example, $\{\equiv\}$. $\{\equiv\}$ = $\{\equiv\}$ and $\{\sqsubseteq\}$. $\{\equiv\}$ = $\{\sqsubseteq\}$.

For computing the composed confidence value, F_{conf} can be designed in different ways (Jung et al., 2009):

- multiplication $F_{conf}(n, n')$ = $n \times n'$
- normalization $F_{conf}(n, n')$ = $(n \times n')/2$, and
- minimization (or maximization) $F_{conf}(n,n')$ = $min(n,n')$ (or $max(n,n')$).

3.3. Comparing the strategies

Using one or other strategy depends on the available resources (intermediary alignments, dictionaries and translators) and features of the languages the ontologies are written:

- **Language features**: if the two languages (L_s and L_t) derive from a same root language (e.g., Latin), they have a similar vocabulary. In such cases, a direct matching can be performed (for instance, directly matching between French and Portuguese ontologies, without applying some multi-lingual strategy). This is not the case when aligning, for instance, Japonese and English ontologies. In such cases, direct translation-based or indirect alignment must be considered, depending on the available resources.

- **Available alignments**: it may be the case that there are no intermediary alignments between the ontologies.
- **Available dictionaries and translators**: it may be the case that available translators or dictionaries do not support the required languages or the domain of the ontologies (e.g. dictionaries provide non-specific terms of a domain and ontologies tend to be domain-specific).

4. API for Multi-lingual Matching

The main contribution of this paper is to describe an API for multi-lingual ontology matching (multi-align API), which specifies the minimal interface for the strategies described above. Figure 1 shows the API class diagram. For sake of brevity, only the signature of the methods of interfaces and abstract classes are shown.

Following the direct translation-based strategy, one source ontology is translated into one translated ontology. In Figure 1, *TranslateOnto* reads the source ontology, translates it, and writes the resulting ontology. The translation is based on a URI translation strategy of labels. *TranslateStrategy* implements *OWLEntityURIConverterStrategy* of OWL-API[4]. For instance, the URI http://www.onto.pt/onto_source.rdf#Tese is converted to http://www.onto.pt/onto_ translated.rdf#Thesis. Such conversion is done using some external resource.

We provide a basic implementation for *TranslateStrategy*, *BasicTranslateStrategy*, which uses the Google-Translator-API[5] to provide the translations. However, Google-Translator can be replaced by a new implementation to be used in *TranslateStrategy*. For *BasicTranslateOnto*, an implementation for *TranslateOnto*, for reading the source ontology and rendering the translated one, we reuse resources provided by the Alignment API 3.6[6](Euzenat, 2004) and OWL-API, respectively. Having the translated and target ontologies, we can use some matcher to match them.

It is important to distinguish between the notions of matcher and strategy. A matcher takes two ontologies and applies some technique to match the entities of these ontologies. For instance, an edit distance (Levenshtein) to match the entities' labels. For the direct translation-based strategy, we apply a step of translation and then the translated and target ontologies are used as input to the matcher. We do not specify a new matcher class in our API because the Alignment API offers a set of different matchers that can be easily reused.

For the indirect strategy, it is assumed that intermediary alignments are generated by some matcher before the composition is performed. We specify the interface *Composition* and provide the corresponding implementation, *BasiComposition*, which reads two ontologies and two alignments (using objects from OWL-API and Alignment API, respectively) and composes a new alignment. This implementation uses a maximisation to compute composed confidences and is restricted to equivalence relations.

The whole package (source, libraries, documentation, and ant file for compilation purposes) can be downloaded at http://www.inrialpes.fr/exmo/people/trojahn/multiapi.

5. Evaluation

5.1. Test Cases

Multi-lingual datasets (ontologies and reference alignments) to evaluate matching approaches are hard to find (there are few databases and some of them are not publicly available). OAEI provides some cases:

- benchmark[7] test 206 (open access): one reference ontology (Test 101) is matched to one French ontology (Test 206). The reference ontology contains 33 named classes, 24 object properties, 40 data properties, 56 named individuals and 20 anonymous individuals.
- very large crosslingual resources[8] (no public access): the thesaurus of the Netherlands Institute for Sound and Vision is matched to two other resources: the English WordNet from Princeton University and DBpedia. A reduced reference alignment is provided for OAEI campaign purposes.
- mldirectory real world case[9] (open access): it consists of matching web sites directories in different languages (English and Japanese). This dataset was not included in 2009 OAEI campaign.

Based on the public test 206 of OAEI, we have created a Portuguese ontology and its corresponding alignments to French and English ontologies (reference alignments). We manually translated the labels of each entity of the ontology 101. The individuals in this ontology are not taken into account.

Such dataset can be downloaded at http://www.inrialpes.fr/exmo/people/trojahn/multiapi.

So, the evaluation is carried out using three test cases:

- EN–FR: English – French ontologies;
- FR–PT: French – Portuguese ontologies;
- EN–PT: English – Portuguese ontologies.

5.2. Matchers

We apply four string-based methods provided by the Alignment API, as matchers:

- *NameEqAlignment*: simple method that compares the equality of ontology entity names and match those objects with the same name;
- *SubsDistNameAlignment*: computes a substring distance on the entity names;
- *EditDistNameAlignment*: uses an editing (Levenshtein) distance between entity names;

[4]http://owlapi.sourceforge.net/
[5]http://code.google.com/p/google-api-translate-java
[6]http://alignapi.gforge.inria.fr/
[7]http://oaei.ontologymatching.org/2009/benchmarks/
[8]http://www.cs.vu.nl/ laurah/oaei/2009/
[9]http://oaei.ontologymatching.org/2008/mldirectory/

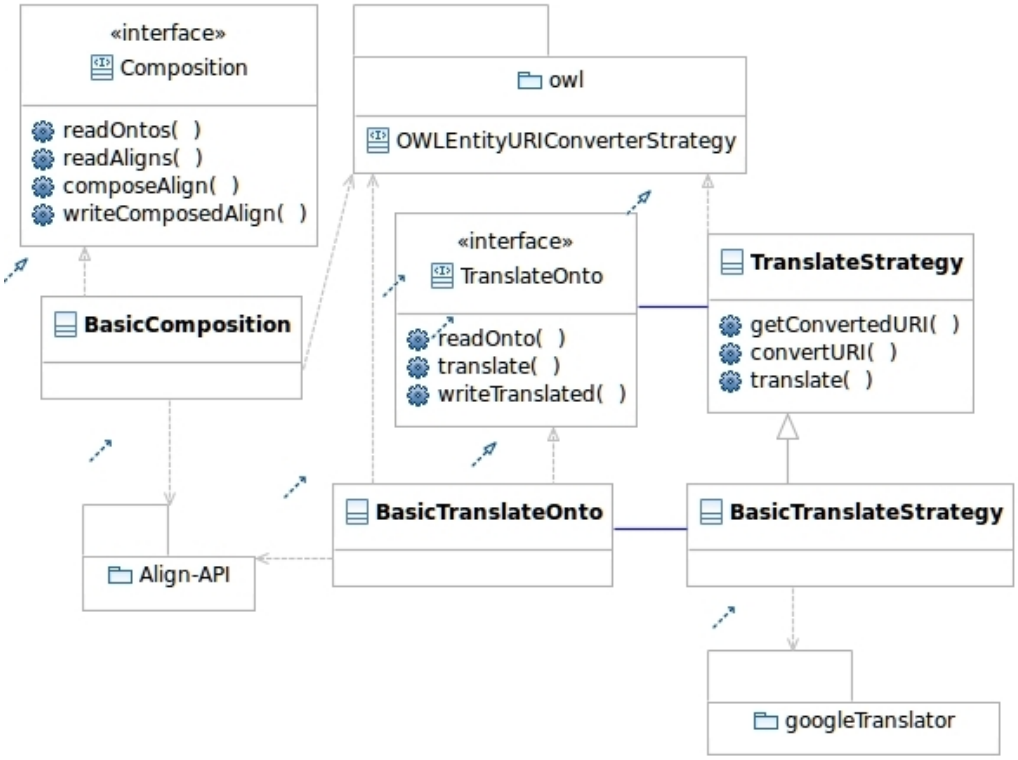

Figure 1: Class diagram for Multi-Align API.

- *SMOANameAlignment*: the similarity between two strings is based on two features: their commonalities and their differences (Stoilos et al., 2005).

Most of the matching systems apply some kind of string-based method. Despite the fact that these methods can be not ideal for matching ontologies written in different languages, they can be seen as a starting point for multi-lingual matching (specially in the cases where the automatic alignment will be used to help users in the matching task).

5.3. Results and discussion

In the experiments, we do not consider the alignments with confidence value lower than 0.55. In the following we present the results for three group of experiments (baseline, direct, and indirect), for each test case.

Table 1 shows the results for the baseline. EditDistName performs better than the other methods in terms of precision, having similar values of recall when compared to SMOAName method. It can be explained by the fact that EditDistName can retrieve alignments such as "Unpublished" and "NonPublie" (FR), which have lower confidence than 0.55 when using the other methods. As expected, NameEq has good precision and low recall.

Table 2 shows the results for the direct translation-based alignment strategy. NameEq performs better than the other methods in terms of precision, while all methods can be considered as having similar values of recall. It can be explained by the way the translations process is carried out, where, for instance, the word "article" (EN) is translated into "item" (PT) (it should be "artigo"), resulting in a false negative. In this way, only the words with very similar strings are retrieved and all methods have similar results in terms of recall.

Table 3 shows the results for the indirect strategy. EditDistName and NameEq have better precision than the other methods, while EditDistName and SMOAName have the best recall. As expected, NameEq returns the same results for all test cases. It is due the fact that only the common names for all three languages are retrieved, for instance, "volume", "isbn", "issn".

Looking for the three groups of experiments, in average, EditDistName and SMOAName can be considered as good matcher candidates.

Regarding each strategy, we must consider the features of each pair of ontologies. Table 4 shows the results for each test case.

For the pair English – French, baseline and indirect present high values of precision while translation-based strategy has high values of recall. We have a similar behaviour for both pairs English – Portuguese and French – Portuguese. However, for the pair English – Portuguese, the translation based strategy improves significantly the recall and the precision is compared with the values of the baseline.

As an expected behaviour, translation-based strategy improves the recall, because the translated terms have higher degree of string similarity than when comparing the original terms. However, one problem associated with this approach is related with the ambiguity in the set of translations (as stated before, we should treat the set of translations instead of retrieving the first item of the list).

In average, looking for F–measure results, translation has

slightly better performance than the baseline and indirect strategy.

6. Concluding Remarks and Future Work

Multi-lingual ontology matching is an important task in ontology matching. This paper has presented an API for multi-lingual matching and a basic implementation. One new test cases was created to show the use of this implementation. These resources are publicly available, representing a starting point for enriching resources in the multi-lingual matching domain, where few contributions have been proposed and resources are increasingly required.

Although our evaluation considers ontologies in one specific domain (in the bibliographic domain the ontologies share similar strings in their terms), we have shown an experiment in which the use of a basic translation approach (which does not consider word sense disambiguation, for instance) surpasses the use of other practical strategies.

The availability of resources such as aligned multi-lingual ontologies is still limited, we hope that this paper inspire further work in this area which we consider as relevant as monolingual alignments and imposes other interesting research questions and challenges.

We intend to address the weaknesses of the approaches in the future. We plan develop further tests using ontologies written in languages that do not have the same root; study how the disambiguation in translations can be performed; enrich the API and improve its implementation, specially taking into account alternative translation resources.

7. References

Jérôme Euzenat and Pavel Shvaiko. 2007. *Ontology matching*. Springer, Heidelberg (DE).

Jérôme Euzenat. 2004. An API for ontology alignment. In *Proceedings of the 3rd International Semantic Web Conference*, pages 698–7112, Hiroshima, Japan.

Jérôme Euzenat. 2008. Algebras of ontology alignment relations. In *Proceedings of the 7th International Conference on The Semantic Web*, pages 387–402, Berlin, Heidelberg. Springer-Verlag.

Bo Fu, Rob Brennan, and Declan O'Sullivan. 2009. Cross-lingual ontology mapping - an investigation of the impact of machine translation. In *Proceedings of the 4th Asian Semantic Web Conference*.

Jason J. Jung, Anne Håkansson, and Ronald Hartung. 2009. Indirect alignment between multilingual ontologies: A case study of korean and swedish ontologies. In *Proceedings of the 3rd KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*, pages 233–241, Berlin, Heidelberg. Springer-Verlag.

Giorgos Stoilos, Giorgos Stamou, and Stefanos Kollias. 2005. A string metric for ontology alignment. In *Proceedings of the 4th International Semantic Web Conference*, pages 624–637, Berlin, Heidelberg. Springer-Verlag.

Cássia Trojahn, Paulo Quaresma, and Renata Vieira. 2008. A framework for multilingual ontology mapping. In *Proceedings of the 6th International Language Resources and Evaluation*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Shenghui Wang, Antoine Isaac, Balthasar A. C. Schopman, Stefan Schlobach, and Lourens van der Meij. 2009. Matching multi-lingual subject vocabularies. In *European Conference on Digital Libraries*, pages 125–137.

Test	refalign			EditDistName			NameEq			SMOAName			SubsDistName		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EN_s - FR_t	1.0	1.0	1.0	0.81	0.48	0.61	0.79	0.20	0.31	0.70	0.52	0.60	0.71	0.38	0.50
EN_s - PT_t	1.0	1.0	1.0	0.76	0.35	0.48	0.70	0.07	0.13	0.63	0.38	0.47	0.68	0.27	0.39
FR_s - PT_t	1.0	1.0	1.0	0.76	0.47	0.58	0.64	0.07	0.13	0.61	0.45	0.51	0.65	0.29	0.40
H-mean	1.0	1.0	1.0	0.78	0.43	0.56	0.73	0.11	0.20	0.65	0.45	0.53	0.68	0.31	0.43

Table 1: Direct alignment.

Test	refalign			EditDistName			NameEq			SMOAName			SubsDistName		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EN_s - FR_t	1.0	1.0	1.0	0.60	0.43	0.50	0.79	0.43	0.55	0.51	0.43	0.46	0.59	0.43	0.49
EN_s - PT_t	1.0	1.0	1.0	0.65	0.51	0.57	0.80	0.51	0.62	0.58	0.51	0.54	0.66	0.51	0.57
FR_s - PT_t	1.0	1.0	1.0	0.45	0.40	0.42	0.55	0.40	0.46	0.41	0.40	0.40	0.47	0.40	0.43
H-mean	1.0	1.0	1.0	0.56	0.44	0.50	0.70	0.44	0.54	0.50	0.44	0.47	0.57	0.44	0.50

Table 2: Direct translation-based alignment.

Test	refalign			EditDistName			NameEq			SMOAName			SubsDistName		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EN_s - FR_t	1.0	1.0	1.0	0.75	0.31	0.44	0.83	0.05	0.10	0.65	0.37	0.47	0.69	0.23	0.34
EN_s - PT_t	1.0	1.0	1.0	0.81	0.31	0.45	0.83	0.05	0.10	0.64	0.30	0.41	0.74	0.21	0.32
FR_s - PT_t	1.0	1.0	1.0	0.86	0.31	0.46	0.83	0.05	0.10	0.75	0.34	0.47	0.75	0.19	0.30
H-mean	1.0	1.0	1.0	0.80	0.31	0.45	0.83	0.05	0.10	0.68	0.34	0.45	0.72	0.21	0.32

Table 3: Indirect alignment.

Test	Baseline			Translation			Indirect		
	P	R	F	P	R	F	P	R	F
EN_s - FR_t	0.75	0.39	0.50	0.62	0.43	0.50	0.73	0.24	0.33
EN_s - PT_t	0.69	0.26	0.36	0.67	0.51	0.57	0.75	0.21	0.32
FR_s - PT_t	0.66	0.32	0.40	0.47	0.40	0.42	0.79	0.22	0.33

Table 4: Baseline, direct translation-based, and indirect results.