Exploiting Multilinguality For Ontology Matching Purposes

Mauro Dragoni

FBK–IRST, Trento, Italy
dragoni@fbk.eu

1 Introduction

The alignment between linguistic artifacts like vocabularies, thesauri, etc., is a task that has attracted considerable attention in recent years [1][2]. With very few exceptions, however, research in this field has primarily focused on the development of monolingual matching algorithms. As more and more artifacts, especially in the Linked Open Data realm, become available in a multilingual fashion, novel matching algorithms are required.

Indeed, in the case of a multilingual environment, there are some peculiarities that can be exploited in order to relax the classic schema matching task:

- the use of multilinguality permits to reduce the problems raised when two different concepts have the same label; indeed, the probability for two diverse concepts to have the same label across several languages is very low;
- multilingual artifacts provide term translations that have already been adapted to the represented domains; therefore, the human creators of a multilingual artifact put a lot of their cultural heritage in choosing the right terms for the each concept.

In this paper, we present a work exploiting the two aspects described above in order to build a multilingual ontology approach for defining mappings between multilingual ontologies. Such an approach, extending the one presented in [3], has been evaluated on domain-specific use cases belonging to the agriculture and medical domains.

2 An Approach for the Matching of Multilingual Thesauri

The proposed approach is based on the exploitation of the labels associated with each concept defined in an ontology. Let us consider two ontologies: (i) a source ontology containing the elements that have to be mapped, and a target ontology used as reference for creating the mappings. The proposed approach has been built by taking inspiration from IR techniques and it exploits the creation of indexes for identifying candidate mappings.

The process is split in two different phases: (i) in the first one, we created the index containing information about the target ontology represented in a structured way; while, (ii) in the second phase, we build queries using information contained in the source ontology for retrieving a rank representing the candidate mappings that we may define between the two thesauri.

Firstly, we extract the whole set of labels from the target ontology and, after a set of preprocessing activities, each concept "C" of the target ontology is transformed into a structured representation containing all multilingual labels describing "C", and all multilingual labels describing concepts belonging to the context of "C" that is the set of concepts directly connected with "C". Such labels are then stored into an index. Then, in the second phase, from each entity of the source index the set of its labels is extracted. A query containing such labels is composed and performed on the index built during the first phase. A rank containing n suggestions ordered by their confidence score is returned by the system and it is used as input for the creation of the mapping that may be done manually from domain experts or automatically by the system.

3 Concluding Remarks

The approach has been evaluated on a set of six multilingual ontologies, coming from the agricultural and medical domains, for which gold standards containing the mappings were available. Then, it has been compared with the previous one presented in [3].

Mapping Set	**# of Mappings**	**Prec. v1**	**Rec. v1**	**F-Measure v1**	**Prec. v2**	**Rec. v2**	**F-Measure v2**
Eurovoc → Agrovoc	1297	0.816	0.874	0.844	0.897	1.000	0.946
Agrovoc → Eurovoc	1297	0.906	0.695	0.787	0.930	0.999	0.963
	Avg.	*0.861*	*0.785*	*0.821*	*0.914*	*1.000*	***0.955***
Gemet → Agrovoc	1179	0.909	0.546	0.682	0.850	0.999	0.918
Agrovoc → Gemet	1179	0.943	0.740	0.829	0.893	0.997	0.942
	Avg.	*0.926*	*0.643*	*0.759*	*0.872*	*0.998*	***0.931***
MDR → MeSH	6061	0.776	0.807	0.791	0.903	0.912	0.907
MeSH → MDR	6061	0.716	0.789	0.751	0.843	0.888	0.865
	Avg.	*0.746*	*0.798*	*0.771*	*0.873*	*0.900*	***0.886***
MDR → SNOMED	19971	0.621	0.559	0.588	0.739	0.826	0.780
SNOMED → MDR	19971	0.556	0.519	0.537	0.871	0.459	0.601
	Avg.	*0.589*	*0.539*	*0.563*	*0.805*	*0.643*	***0.715***
MeSH → SNOMED	26634	0.690	0.660	0.675	0.741	0.814	0.776
SNOMED → MeSH	26634	0.657	0.564	0.607	0.831	0.544	0.658
	Avg.	*0.674*	*0.612*	*0.642*	*0.786*	*0.679*	***0.729***

Table 1: Comparison between the results obtained by the previous version of the system and the proposed one.

References

1. Euzenat, J., Shvaiko, P.: Ontology matching. Springer (2007)
2. Bellahsene, Z., Bonifati, A., Rahm, E., eds.: Schema Matching and Mapping. Springer (2011)
3. Dragoni, M.: Exploiting multilinguality for creating mappings between thesauri. In: Proceedings of the 30th Annual ACM Symposium on Applied Computing. SAC 2015, ACM (2015) 382–387