A Term-Based Approach for Matching Multilingual Thesauri

Mauro Dragoni[2], Andi Rexha[3], Matteo Casu[1], and Alessio Bosca[1]

[1] Celi s.r.l., Via S.Quintino 31, I-10131, Torino, Italy
[2] FBK–IRST, Trento, Italy
[3] Know-Center Graz, Graz, Austria
`dragoni@fbk.eu arexha@know-center.at casu|alessio.bosca@celi.it`

Abstract. In this paper, we present a multilingual matching approach aiming at building matches between terms belonging to multilingual thesauri. The approach is presented as a variant of the schema matching problem and present its evaluation on domain-specific use cases by demonstrating the viability of the proposed technique for facing the multilingual thesaurus matching approach.

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 term-based approach for defining mappings between multilingual thesauri. Such an approach 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 term defined in a thesaurus. Let us consider two thesauri: (i) a source thesaurus containing the elements that have to be mapped, and a target thesaurus used as reference for

creating the mappings. The proposed approach has been built by taking inspiration from information retrieval techniques and it exploits the creation of indexes for identifying candidate mappings.

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

First of all, the two thesauri are considered with two different roles: a source thesaurus that is used as starting point for the creation of the mapping, and a target thesaurus that is considered as ending point of the mapping. It is split in two main phases: in the first one, it operates on the target thesaurus, while in the second one, on the source thesaurus. Firstly, we extract the whole set of labels from the target thesaurus and, after a set of preprocessing activities, each term of the target thesaurus is transformed into a structured representation containing all its multilingual labels and it is 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 thesauri for which gold standards containing the mappings were available. Such thesauri belong to two different domains: three thesauri to the agricultural and environment domain, while the other three to the medical one. The promising results shown in Table 1 demonstrated the effectiveness of the proposed approach.

Mapping Set	# of Mappings	Prec@1	Prec@3	Prec@5	Recall
Eurovoc → Agrovoc	1297	0.861	0.946	0.978	0.785
Gemet → Agrovoc	1181	0.927	0.973	0.988	0.643
MDR → MeSH	6061	0.746	0.901	0.948	0.799
MDR → SNOMED	19971	0.589	0.793	0.882	0.539
MeSH → SNOMED	26634	0.674	0.853	0.920	0.612

Table 1: Results obtained on the multilingual ontologies used for the Context 1 evaluation.

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. Choi, N., Song, I.Y., Han, H.: A survey on ontology mapping. SIGMOD Rec. **35**(3) (September 2006) 34–41