PMKI: an European Commission action for the interoperability, maintainability and sustainability of Language Resources

Peter Schmitz, Enrico Francesconi*, Najeh Hajlaoui, Brahim Batouche
Publications Office of the European Union, Luxembourg
* Publications Office of the European Union, Luxembourg and ITTIG-CNR, Florence, Italy
{Peter.Schmitz, Enrico.Francesconi}@publications.europa.eu
{Najeh.Hajlaoui, Brahim.Batouche}@ext.publications.europa.eu

Abstract

The paper presents the Public Multilingual Knowledge Management Infrastructure (PMKI) action launched by the European Commission (EC) to promote the Digital Single Market in the European Union (EU). PMKI aims to share maintainable and sustainable Language Resources making them interoperable in order to support language technology industry, and public administrations, with multilingual tools able to improve cross border accessibility of digital services. The paper focuses on the main feature (interoperability) that represents the specificity of PMKI platform distinguishing it from other existing frameworks. In particular it aims to create a set of tools and facilities, based on Semantic Web technologies, to establish semantic interoperability between multilingual lexicons. Such task requires to harmonize in general multilingual language resources using standardised representation with respect to a defined core data model under an adequate architecture. A comparative study among the main data models for representing lexicons and recommendations for the PMKI service was required as well. Moreover, synergies with other programs of the EU institutions, as far as systems interoperability and Machine Translation (MT) solutions, are foreseen. For instance some interactions are foreseen between PMKI and MT service provided by the EC but also with other NLP applications.

Keywords: Language Resources, Interoperability, Maintainability, Sustainability.

1. Introduction

Language barriers in the EU make the European market fragmented and decrease its economic potential. The EU institutions aim to overcome language obstacles and increase cross-border e-commerce by building open multilingual tools and features free of charge. For this reason the European Commission, through the ISA^2 program[1], launched a pilot project on creating a public multilingual knowledge management infrastructure aimed to support e-commerce solutions in a multilingual environment. By creating interoperable multilingual classifications and terminologies, easily reusable by small and medium-sized enterprises (SMEs) and public administrations, the project aims to support services like machine translation, localisation and multilingual search. SMEs are currently at a disadvantage compared to big companies due to the high cost of providing multilingual services. In this respect PMKI aims to create a set of tools and facilities, based on semantic Web technologies, aimed to support enterprises, in particular the language technology industry, with multilingual tools in order to improve cross border accessibility of digital services and e-commerce solutions.

In practical terms, overcoming language barriers on the Web means creating multilingual lexicons (as vocabularies, thesauri, taxonomies, semantic networks), establishing links between concepts, as well as using them to support the accessibility of services and goods offered through the Internet.

This paper is organized as follows: in Section 2 the main objectives of PMKI are discussed; in Section 3 the foreseen interoperability solutions are illustrated; finally in Section 4 some conclusions are reported.

2. PMKI

The objective of PMKI is to implement a proof-of-concept infrastructure to expose and to harmonize internal (European Union institutional) and external multilingual lexicons aligning them in order to facilitate interoperability. Additionally the project aims to create a governance structure to extend systematically the infrastructure by the integration of supplementary public multilingual taxonomies/terminologies.

PMKI is a pilot project to check the feasibility and to prepare a road map to convert such proof of concept into a public service.

The need to have a public and multilingual platform that can play the role of a hub to collect and to share language resources in standardised formats is essential to guarantee semantic interoperability of digital services. For instance, such platform is missing in CEF.AT[2], while it would provide an advantage for the development of machine translation systems. In particular it can provide alignments of domain specific terminologies for developing specific-domain translation systems (tender terminology, medical terminology, etc.).

A platform like PMKI may represent a one-stop-shop harmonized multilingual lexicons repository at European level.

Complementary to the European Language Resource Coordination (ELRC[3]) action, which aims at identifying

[1] ISA$^{2:}$ Interoperability solutions for public administrations, businesses and citizens (https://ec.europa.eu/isa2/home_en)

[2] https://ec.europa.eu/digital-single-market/en/automated-translation

[3] ELRC: the European Language Resource Coordination action launched by the European Commission as part of the CEF.AT platform activities, to identify and gather language data across all 30 European countries participating in the CEF programme. This will be followed up by actions concerning the setting up of a repository of language resources for CEF AT and further

and gathering language and translation data, PMKI platform aims first to harmonize multilingual language resources making them interoperable, then to integrate supplementary public multilingual taxonomies/terminologies in a standardized representation. That is why we need first to define 1) a sophisticated standard representation that will be used with respect to 2) a defined core data model (in case with extensions) under 3) an adequate architecture.

These three requirements were respectively analysed and detailed in three first analysis phases of the project:

- Analysis of existing relevant standards for the representation of lexicons that will be made available on the PMKI platform;
- PMKI core data model and extensions (based on the standard representation that was recommended as result of the previous analysis);
- Analysis of available platforms for managing lexicons.

Interoperability is one of the main features of the PMKI platform; such platform will provide support to develop multilingual tools such as machine translation, localization, search etc. For instance for the machine translation tool, interoperable translation data is a factor of success to improve the quality mainly for under-resourced languages.

3. Interoperability

Interoperability between language resources is essential in order to facilitate the access to European service by overcoming the language barriers. In PMKI we started to apply such interoperability by aligning language resources.

3.1 Experimental approach

The alignment idea is based on the work of Bartoloni and Francesconi (2010). They proposed a framework based on the definition of an isomorphism between linguistic resources alignment problem and the Information Retrieval (IR) problem.

To have a linguistic resource amenable for computation, we first organize the linguistic resource according to a specific standard. Then in order to create a linguistic resource alignment system we follow the following processing steps that (Figure 1):

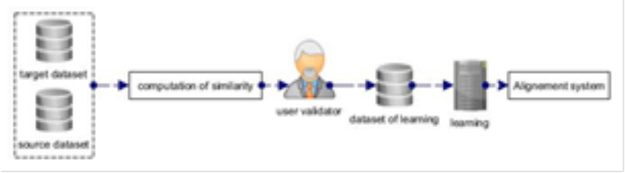

Figure 1: Steps of alignment with supervised learning.

1. Give a representation of the semantics of linguistic resources amenable for computation.
2. Compute similarities between source and target resources to create dataset for learning.
3. Validate the matching of linguistic resource with user validator to build supervised learning dataset.
4. Launch the learning process to create an alignment system.

As EuroVoc[4] is the kernel of PMKI platform, we use it as source dataset to be aligned with new target dataset. All resources are indexed by their URI. The storage of new resource follows SKOS, LEMON, or ONTOLEX standards.

The diversification of linguistic resources is a necessary condition to have a good coverage of the learning phase, able to deal with different written/pronounced word and multi-word expressions. This diversification problem is one of the sensitive aspects in PMKI.

We started collecting linguistic datasets, each one represented as a graph of triples. Each dataset is represented as one of the three standards mentioned above, each standard has some useful properties for alignment. We selected the value of those properties using SPARQL query and through the SPARQL endpoint of the related triple store. Therefore, those selecting property values are used to instantiate the resources (according to Section 3.3). We used information retrieval tool (Lucene[5] as java API) to create and to represent the index of terms for our alignment approach.

3.2 Formalization of alignments

Our matching methods are string-based, language-based, and constraint-based techniques. From these three techniques, the string-based matching technique allows to represent the input format for the alignment. The language-based technique is useful to extract useful words for the string-based technique. Therefore we use an hybridization of matching techniques where language-based matching is used before the string-based one.

3.3 Linguistic resource formalization

Mapping between linguistic resources aims at matching terms semantics rather than searching for lexical equivalences. In traditional linguistic resources descriptors and non-descriptors are represented by different terms (such as `lemon:writtenRep`, `lemon:usage`, `skos:prefLabel` and `skos:altLabel`) expressing the same meaning. More precisely, each meaning is expressed by one or more terms in the same language (e.g. *pollution, contamination, discharge of pollutants*), as well as in different languages. Moreover each term can have more than one sense, i.e. it can express more than one concept. Therefore to effectively map linguistic resource, term (simple or complex) semantics has to be captured and represented.

In Information retrieval (IR) a query is usually constructed as a context (set of keywords) able to better represent the semantics of the users' information needs. Similarly in linguistic resource mapping (LM) the semantics of a term (simple or complex) is conveyed not

data collection and awareness actions in the context of calls for tenders and calls for proposals for which the selection procedure is still ongoing. More information can be found here: http://www.lr-coordination.eu/

[4] http://eurovoc.europa.eu/4067

[5] https://lucene.apache.org/

only by the keywords, but also by the context in which the term is used as well as by the relations with other terms. In the LM problem, we aimed at identifying logical views and related framework for linguistic resource concepts representation able to better capture the semantics of terms in source and target linguistic resource, as well as to measure their conceptual similarity.

Bartoloni and Francesconi (2010) proposed to represent the semantics of a linguistic resource by a vector *d* of binary entries composed by the term itself, relevant terms in its definition, in the alternative labels, as well as terms of directly related linguistic resource (`skos:broader`, `skos:narrower`, `skos:related` concepts).

A vocabulary of normalized terms from target linguistic resource is constructed, where "normalization" in this context means string pre-processing, in particular word stemming and stop-words elimination. Being *T* the dimension of such vocabulary, both source and target linguistic resource concepts are represented in a vector space of *T*-dimension ($d=[x_1,x_2,...x_T]$); the entry x_i gives information on the presence/absence of the corresponding x^{th} vocabulary term among the terms characterizing the linguistic resource *d*. In (Figure 2) a binary vector representation of EuroVoc concept is sketched.

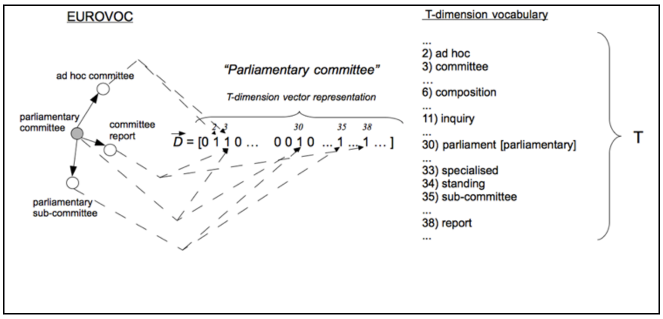

Figure 2: T-dimension vectorial representation of a thesaural description *d* ***d***

3.4 Formal characteristic of linguistic resource mapping

Such kind of characteristics allow us to propose a definition of the schema-based Linguistic-resource Mapping (LM) problem as an Information Retrieval (IR) problem: the aim is to find concepts in a target linguistic resource, which match with the semantics of a given source concept. The isomorphism between LM and IR can be established once we consider a source language resource as a query of the IR problem, and a target language resource as a document (Baeza-Yates, 1999).

Therefore the LM problem can be viewed and formalized, as an IR problem composed of a 4-uple *LM = [D, Q, F, R(q,d)]* $LM{=}D,Q,F,Rq,d$ where:

1. *D* $\boldsymbol{D}$is a set of the possible representations (logical views) of a concept in the target linguistic resource (i.e. a document to be retrieved in the IR problem);

2. *Q* $\boldsymbol{Q}$is the set of the possible representations (logical views) of a concept in the source linguistic resource (a query in the IR$IR$ problem);

3. *F* $\boldsymbol{F}$is the framework of resource representation in source and target linguistic resource (in our case the vectorial space of T-dimension endowed with metrics);

4. *R(q,d)*$\boldsymbol{Rq,d}$ is a ranking function, which associates a real number with *(q,d)* where $q \in Q$ $and\ d \in D$ giving an order of relevance to the resource in the target concept with respect to the one of the source.

In this framework the implementation of a linguistic resource mapping procedure is represented by the instantiation of the four previous components.

3.5 Evaluation of similarity

The similarity is inversely proportional to the distance between binary vectors representing the linguistic resources. Therefore, the function has the input parameters *q*and *d* which are the vectors detailed previously. Having represented the semantics of linguistic resource as a binary vector, their similarity can be measured as the related binary vectors correlation, quantified, for instance, as the cosine of the angle between them.

$Sim(q,d)= q \times d / |q| * |d|$

Where |*q*| and |*d*| are respectively the norms of the vectors representing concepts in the source and in the target linguistic resource. In our case, a source concept will belong to EuroVoc data set, and the target of the linguistic resource will belong to any linguistic dataset to be aligned. The size of vectors *q* and *d* corresponds to the size of the list of dictionary words; each word has a position in the vector representing a concept. A binary value in the vector is equal to 1 if the word of that position is founded in the description of the concept; otherwise it is equal to 0 as shown in figure 2.

3.6 Supervised Learning

Having established a proper similarity measure between concepts, a criterion able to predict matching concepts has to be defined. In Francesconi et al (2008) a criterion was implemented by defining a heuristic threshold over a similarity measure: if the similarity between linguistic resource is over a threshold, a `skos:exactMatch` relation is established. Such strategy, anyway, usually suffers from generalization capabilities out of the matching examples used to tune the heuristics. Generalization capabilities for a prediction strategy can be introduced by adopting machine learning techniques able to learn a predictive function from a training set of matching relations. In Bartoloni and Francesconi (2010) such predictive function is obtained by a Support Vector Machine (SVM)[6] trained to classify a pair of descriptors into two classes *{Match (+1), no-Match (-1)}.*

[6] In order to select the best technique, other machine learning algorithms will be compared to SVM such as *Multi-Layer Perceptron* (MLP).

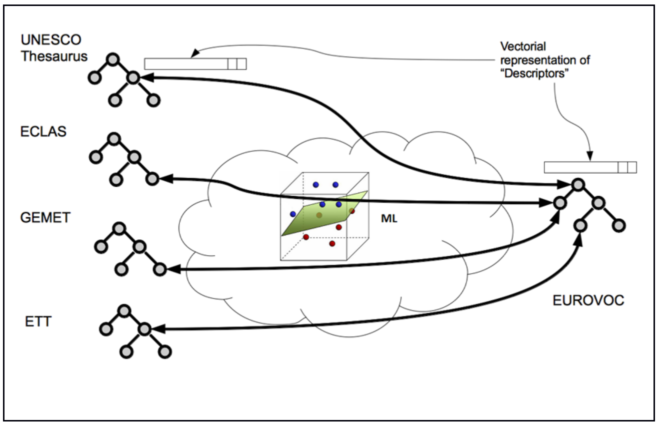

Figure 3: Mapping of linguistic resource using Machine Learning (ML)

A training set for the SVM linguistic resource matching predictor is composed by training examples described by vectors of features deemed representative for descriptors conceptual matching: in particular the i^{th}example is represented by a feature vector φ_i associated to a pair *(d_l,q)* of a source and target linguistic resource respectively, including:

- the similarity measure s*im(d_l,q)*, computed according to the cosine function;
- the logical view of the target descriptor d_l

together with a relevance judgment *y={+1,-1}* for that target descriptor (d_l) on that source descriptor *(q)* that is either matching (+1) or non-matching concept (-1). Therefore a generic i^{th} training example describing a pair of linguistic resource descriptors and related relevance judgement is

$$\varphi_i=<< sim\ (d_l,q),\ d_l>,y>$$

On the basis of such training set[7], the goal is to build a classifier (a separating surface) which is able to distinguish between matching and non-matching descriptors. The classifier provides also the distance of the examples from the separating surface, giving a measure of the prediction confidence, thus allowing a ranking among candidate target descriptors. The best ranked descriptor is finally chosen as the predicted matching concept.

4. Conclusion and perspectives

This paper presents the preliminary results obtained within the PMKI project for implementing interoperability solution of lexical resources. In particular the main Semantic Web standards available in literature for representing lexicons have been identified and their characteristics were analysed. For the ability of describing lexical components in different languages using LEMON, the related concepts and their mapping relations using SKOS, the Ontolex standard resulted as the preferred model to be adopted as reference for the PMKI platform.

In this work, we formalized the alignment problem as information retrieval problem that can be treated using supervised machine learning techniques. Different data tests are required to train the machine learning algorithm. Our approach can be summarised into three points: 1) Knowledge representation of linguistic standards (SKOS, LEMON and ONTOLEX), 2) Processing framework and steps of alignment and 3) Formalization of linguistic resource, dataset of learning, similarity functions, supervised learning.

The next phase of the project will provide an evaluation of mapping algorithms and propose a technical infrastructure for the implementation and maintenance of lexical resources and their interoperability.

5. Acknowledgements

This work has been drafted under the 2016.16 action[8], which is supported by the ISA² Programme[9]. ISA² is a EUR 131 million programme of the European Commission, supporting the modernisation of public administrations in Europe through the development of eGovernment solutions. More than 20 solutions are already available[10], with more to come soon. All solutions are open source and available free of charge to any interested public administration in Europe.

6. Bibliographical References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. *Addison Wesley*.

Bartoloni, G. and Francesconi, E. (2010). Sharing Knowledge by Conceptual Mapping: the case of EU Thesaural Interoperability. In Proceedings of the *JURIX Conference*, pp. 17-26, ISBN 978-1-60750-681-2, IOS Press, 2010, DOI 10.3233/978-1-60750-682-9-17.

Bosque-Gil & al., (2015) Applying the OntoLex Model to a Multilingual Terminological Resource, *ESWC* 2015

Francesconi, E. Faro, S. and Marinai, E. (2008). Thesauri alignment for eu egovernment services: a methodological framework. In Proceedings of the *JURIX 2008 Conference*, pp. 73-77, IOS Press.

Euzenat, J. Shvaiko, P. (1998). Ontology Matching, Springer-Verlag, Berlin Heidelberg (DE), 2013xviii+511pp. (incl. bibl., index, exercises and solutions), 98 figures, 20 tables, 646 references, ACM Classification: H.3, H.4, I.2, F.4.

Miles, A. Matthews, B. Wilson. M, and Brickley. D. (2005) SKOS Core: Simple knowledge organisation for the Web, DCMI.

Villegas, M. & Bel, N. (2015). PAROLE/SIMPLE 'lemon' ontology and lexicons. *Semantic Web*, 6, 363-369.

[7] The training set is built by human experts from a set of linguistic resources of interest for the European institutions.

[8] https://ec.europa.eu/isa2/actions/overcoming-language-barriers_en

[9] https://ec.europa.eu/isa2/

[10] https://ec.europa.eu/isa2/solutions_en