BUILDING MULTILINGUAL LEXICAL RESOURCES ON SEMIOTIC PRINCIPLES

Davide Picca
Center of Computational Learning Systems, Columbia University, 475 Riverside Drive, New York City, U.S.A.

Keywords: Upper ontology, Multilingual Resources, Knowledge representation, Applications and case-studies.

Abstract: Multilinguality permeates the web so that multilingual resources are fundamental in several NLP applications as cross language information retrieval as well as machine translation. Nonetheless the manual creation of such resources is very expensive. Semantic Web technologies can represent a great enhancement for NLP applications. In this paper, we show how Semantic Web technologies as an upper ontology based on well-founded semiotic theories can be applied to build multilingual lexical resources as Machine Readable Dictionaries (MRDs).

1 INTRODUCTION

In computer science, semantics, along with a conceptual ontological structure, can really enlarge the vision to new applications and achieve new goals in multilingual machine translation and multilingual knowledge management. Providing Linking Open Data Community Project (W3C, 2009) with an underlying theoretical model can deeply enhance the use of these data for larger projects reducing the complexity of queries and computational times. This paper aims at providing this kind of structure in order to offer a solid and well-founded ontological structure for genuine multilingual dictionaries. As far as we know, there are not attempts to build multilingual dictionaries organized on the base of semiotic laws. Our approach takes advantage of well-founded linguistic theories in order to build machine readable dictionaries leveraging on external resources such as DBpedia and semantic web standards such as OWL. The Linguistic Meta-Model (LMM) (Picca et al., 2008) ontology provides a solid semiotic-oriented support for multilingual linguistic knowledge by means of which we are able to fully exploit multilingual semantic web datasets as DBpedia in order to build solid and well-structured multilingual MRDs as shown in Section 4 and the intent of the semantic web to enhance the web reorganizing data into an out-and-out knowledge repositories (Jain et al., 2010) can be achieved.

2 RELATED WORK

Creation of multilingual lexical resources are widely explored in both domains, NLP and Semantic Web. In the former, investigation of automatic methods for creating MRDs has a old-established tradition (Utsuro et al., 1994). Many scholars have faced the issue of machine readable dictionary and the literature is vast (Fung, 1998). Three main approaches can be outlined: the context based approach (Morin et al., 2007) and syntactical analysis approach (Yu and Tsujii, 2009). A lexical repository based on theoretical foundations is WordNet and its multilingual extensions EuroWordNet (Vossen, 2002) and MultiWordNet (Bentivogli et al., 2002). In the Semantic Web field, specific relations between individual ontologies and lexica are addressed in literature quite often, e.g. (Gangemi et al., 2002) and recent approach comes from the semantic web field and it is described in (Auer et al., 2007).

3 A FORMAL LEXICAL META MODEL

Semiotic principles (Peirce, 1958) allow to abstract from individual lexical standards, by providing a semiotic interface between specific semantics of different lexica. The appropriateness of this kind of meta-model is that any interface or translation method

Picca D..
BUILDING MULTILINGUAL LEXICAL RESOURCES ON SEMIOTIC PRINCIPLES.
DOI: 10.5220/0003099504120415
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2010), pages 412-415
ISBN: 978-989-8425-29-4

can refer to a unique façade. The communication laws underlying semiotics are particularly suitable for computer science applications. In fact, such as discipline tries to highlight structures governing communication processes either formal or informal making them easier to compute. LMM ontology is composed of three main classes: `Reference`, `Meaning` and `Expression`. (See (Picca et al., 2008) for further details)

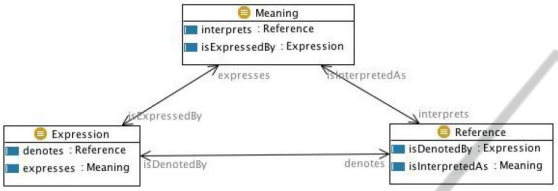

Figure 1: The semiotic triangle in our ontology.

The `Reference` level, represented by Figure 2, is populated by any possible individual in the logical world, being it either a concrete object or any other social object whose existence is stipulated by a community. Individuals are related by the fact that they co-occur into events.
Instances of the class `Reference` are all those *entities* belonging to the universe of discourse, including e.g. *physical objects*, *events*, etc., and they have an explicit reference "in the world".

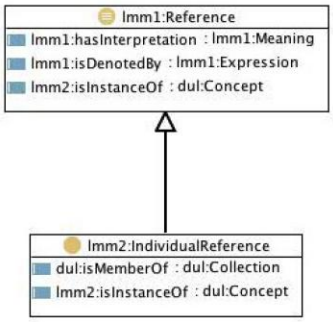

Figure 2: The class `Reference`.

Concepts are represented as instances of `Meaning` objects. Concepts are related between each other in two different ways. Subsumption relations organize concepts into hierarchies of subclasses (e.g. the dog is an animal). These relations are reflected at the extensional level by the fact that the set of instances denoted by the subclasses are contained into the set of instances of their superclasses. Conceptual relations are in turn represented by descriptions (whose definition is mutuated from the Description and Situations framework (Gangemi et al., 2002)), expressing the possibility for events to occur. Descriptions can be considered reified relations, therefore they are actually instances of the meaning class themselves. All those classes are explicitly inherited from the *Descriptions and Situations framework* as represented in DOLCE-Ultralite (Gangemi et al., 2002), and they aim at catching the basic semiotic aspects involved in semantic technologies. Thanks to LOD project (Bizer and Heath, 2007) and the LMM ontology, users are able to leverage on well founded semiotic theories in order to build more complete Machine Readable Dictionaries. `Expression`. Finally, the two layers of meaning and reference are connected to the language by means of the `Expression` layer (see Figure 3). Expressions are social objects produced by agents in the context of communicative acts. They are natural language terms, symbols in formal languages, icons, and whatever can be used as a vehicle for communication. Expressions denoting concepts, frames and topics (such as person, drink and sport) are interpreted by means of their connections to the meaning layer, while expressions denoting instances are directly connected to their corresponding individuals.

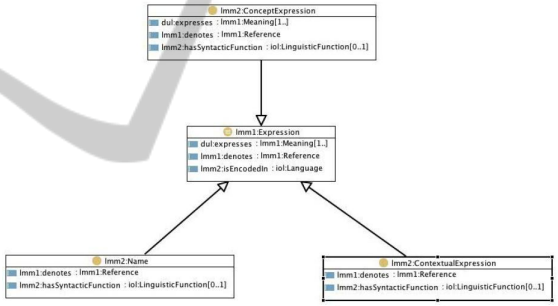

Figure 3: The class `Expression`.

4 METHOD AND IMPLEMENTATION

DBpedia is a semantic database built using structured information from Wikipedia. Its pages are designated by unique string identifiers and, as underlined in (Kazama and Torisawa, 2007), there are different parts in the structure of an article that can be isolated using the syntax of the source files and used for knowledge extraction. One of the nicest feature of Wikipedia is the inter-language linkage. An inter-language link is a direct connection between two articles in different languages. In order to extract multilingual terminology, we use the inter-language links.

The method of extraction is explained in detail as follows. Let t be the source term to be translated. Let a_t be the corresponding DBpedia article for the term

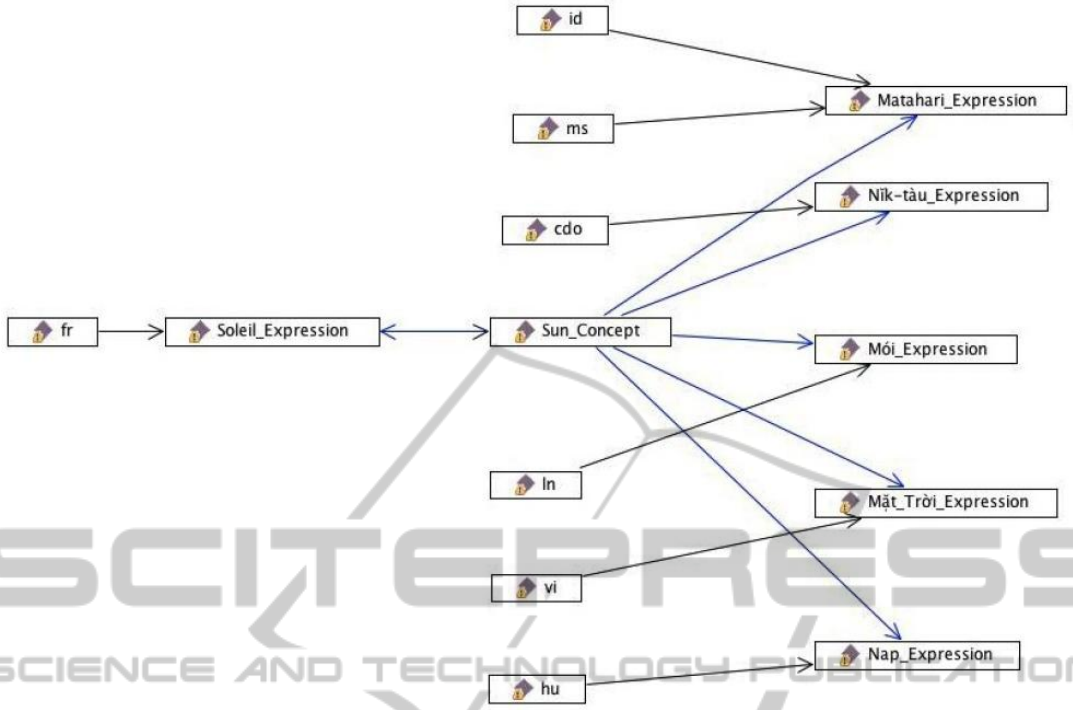

Figure 4: The ontology allows to retrieve all lexical forms of a given concept at a glance.

t. Let il^{l}_{a} be the inter-language link from the article *a* to the corresponding target article in the language *l*. So the set of translation candidates is defined as $TC = \sum_{l=1}^{L} il^{l}_{a}$. In addition for each term *t*, we also extract DBpedia categories to which the term belong and the first paragraph *g* of the Wikipedia article as a definition of the term itself. So, we have one additional set: $C = \{c_1, ..., c_n\}$ if $t \in c$ where c is the category to which the term belong and one additional information *g*.

Once the set T is built, we perform an automatic mapping between *t* and each element in *TC*. The mapping aims to give a structure to each item within the ontological formalization.

We finally added some properties in order to link all this information together and provide a coherent and easy-to-access structure. In particular we added the following ontological properties:

- *isDenotedBy* linking a `Reference` and an `Expression`
- *hasInterpretation* linking a `Reference` and a `Description`
- *isRepresentationLanguageOf* linking a `NaturalLanguagage` and an `Expression`
- *isTopicOf* linking a `Topic` and a `Concept`
- *defines* linking a `Description` and a `Concept`
- *isSubOrdinatedTo* linking a `Concept` and a `Concept`
- *expresses* linking an `Expression` and a `Concept`
- *hasInterpretant* linking an `Expression` and a `Concept`

5 SOME EXAMPLES

For sake of simplicity and organization, we limited the examples of this paper to the concept *Sun* . Queries are more human intuitive and close to human natural language and they can be conceived using very generic semiotic properties easy to implement for complex tasks. Queries based on semiotic principles allows to overcome the specific linguistic structure (syntax or orthography for example) of a language gathering under a unique framework different specific structures. Such as structures would be incomparable without a conceptual higher level (see Section 3). DBpedia does not allow this kind of queries since it does not provide a genuine linguistic structure. The distinction between `Concept`, `Reference` and `Expression` has the capability of highlighting the deeper structure of linguistic items making queries simpler and more human cognitive oriented.

Table 1: The SPARQL query using a semiotic-oriented upper ontology.

SELECT ?*Expression* WHERE { :**Sun_Concept** `dul:isExpressedBy` ?*Expression* ?*Language* `dul:IsEncodedIn` :**it**}

The same runs to retrieve lexical forms. Using a simple SPARQL query we get several lexical forms for the concepts *Sun* as depicted in Figure 4

Another interesting feature is the capability of gathering all topics of a concept in synoptic view. By using the class `Topic`, it is possible to put together all categories to which the concept belongs.

Multilingual dictionaries often are limited to provide just the translations of terms without giving the gloss. Glosses help to define the concept supporting the sense understanding. WordNet (Miller, 1990) is a good example of the use of glosses to define the sense of a term. The class `Description` aims at providing this additional information exploiting the first paragraph of the Wikipedia article as a good candidate to describe the concept.

6 CONCLUSIONS

In this paper we presented a novel technique to create multilingual resources from DBpedia within a semiotic-oriented upper ontology framework. We showed limitations and potentialities of Dbpedia to be used as a terminological repository. In order to fully exploit this resource, we created a powerful ontological structure based on well-founded semiotic theories.

In the future works, we plan to add new features as the management of linguistic forms as discussed in Section 1 and we plan to use it as support for multilingual machine translation. We aim at creating a multilingual system able to translate from many-to-many language using an ontological support.

ACKNOWLEDGEMENTS

This research has been funded by the Swiss National Science Foundation (SNSF) for Project no. 127651 of the Individual Support Fellowship program. Moreover, I would like to thank Aldo Gangemi from the Laboratory for Applied Ontology to share with me some ideas concerning the theoretical foundations underlying this paper.

REFERENCES

Auer, S., Bizer, C., Kobilarov, G., and Lehmann, J. (2007). Dbpedia: A nucleus for a web of open data. *In proceedings of 6th Int'l Semantic Web Conference (ISWC2007)*, pages 722–735.

Bentivogli, L., Pianta, E., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

Bizer, C. and Heath, T. (2007). Interlinking open data on the web. In *Proceedings Poster Track, Extended Semantic Web Conference*.

Fung, P. (1998). A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *AMTA*, pages 1–17.

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2002). Sweetening ontologies with dolce. In *Proceedings of European Workshop on Knowledge Acquisition, Modeling and Management (EKAW)*, pages 166–181.

Jain, P., Hitzler, P., Yehy, P. Z., Vermay, K., and Shet, A. P. (2010). Linked data is merely more data. In *Linked Data Meets Artificial Intelligence*, pages 82–86.

Kazama, J. and Torisawa, K. (2007). Exploiting wikipedia as external knowledge for named entity recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707.

Miller, G. A. (1990). Nouns in wordnet: a lexical inheritance system,. *International Journal of Lexicography*, 3(4):245–264.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *ACL*.

Peirce, C. S. (1958). *Collected Papers of Charles Sanders Peirce*. MIT Press, Cambridge, Mass.

Picca, D., Gliozzo, A. M., and Gangemi, A. (2008). Lmm: an owl-dl metamodel to represent heterogeneous lexical knowledge. In *LREC*.

Utsuro, T., Ikeda, H., Yamane, M., Matsumoto, Y., and Nagao, M. (1994). Bilingual text, matching using bilingual dictionary and statistics. In *Proceedings of the 15th conference on Computational linguistics*, pages 1076–1082, Morristown, NJ, USA. Association for Computational Linguistics.

Vossen, P. (2002). Eurowordnet: general document. Technical report.

W3C, C. (2009). Linkingopendata. http://bit.ly/dW03.

Yu, K. and Tsujii, J. (2009). Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 121–124, Morristown, NJ, USA. Association for Computational Linguistics.