OntoEnrich: A Platform for the Lexical Analysis of Ontologies

Manuel Quesada-Martínez[1(✉)], Jesualdo Tomás Fernández-Breis[1], Robert Stevens[2], and Nathalie Aussenac-Gilles[3]

[1] Facultad de Informática, Universidad de Murcia, IMIB-Arrixaca, CP 30100 Murcia, Spain
{manuel.quesada,jfernand}@um.es
[2] University of Manchester, Oxford Road, Manchester M13 9PL, UK
stevens@cs.manchester.ac.uk
[3] Université Paul Sabatier, IRIT, 118 Route de Narbonne, F-31062 Toulouse, France
aussenac@irit.fr

Abstract. The content of the labels in ontologies is usually considered hidden semantics, because the domain knowledge of such labels is not available as logical axioms in the ontology. The use of systematic naming conventions as best practice for the design of the content of the labels generates labels with structural regularities, namely, lexical regularities. The structure and content of such regularities can help ontology engineers to increase the amount of machine-friendly content in ontologies, that is, to increase the number of logical axioms.

In this paper we present a web platform based on the OntoEnrich framework, which detects and analyzes lexical regularities, providing a series of useful insights about the structure and content of the labels, which can be helpful for the study of the engineering of the ontologies and their axiomatic enrichment. Here, we describe its software architecture, and how it can be used for analyzing the labels of ontologies, which will be illustrated with some examples from our research studies.

1 Introduction

Many ontologies have been developed in recent years. For instance, BioPortal (http://bioportal.bioontology.org/) contains 388 biomedical ontologies at the time of this writing. Ontology authors include strings of characters as labels that describe ontology classes. These labels can embed hidden semantics that is not represented as logical axioms in the ontology. The Open Biomedical Ontologies (OBO) Foundry defines criteria to be followed by biomedical ontology authors such as the use of a systematic naming in ontology labels. Then, the analysis of regularities in ontology labels might help to detect hidden semantics. For example, in the Gene Ontology Molecular Function ontology (GO-MF) [2], regularities like "binding" can be converted into patterns like "X binding" that enrich the ontology with axioms like *"subClassOf enables some (binds some ?x)"*; and these axioms can be re-used in other more specific patterns like "X receptor binding" and "X domain binding".

P. Lambrix et al. (Eds.): EKAW 2014 Satellite Events, LNAI 8982, pp. 172–176, 2015.
DOI: 10.1007/978-3-319-17966-7_25

Our hypothesis is that supporting ontology authors in the analysis of the lexical regularities (LRs) from ontology labels can result in axiomatically enriched ontologies, which should be more useful for their application in real projects. Tools like OntoCheck and OntoCure (http://protegewiki.stanford.edu/wiki/OntoCheck) foster lexical harmonization in ontology labels. Besides, Caméléon [1] uses a supervised process of candidate patterns for relation acquisition from texts, and they use and refine patterns from a catalog. In this paper, we present the OntoEnrich platform, which implements our method for lexical analysis (LAs) and characterization of ontologies [3]. The main advantage of our method is providing tools for the interactive analysis of LRs, which could elucidate patterns like the Caméléon candidate patterns but without starting from a catalog.

2 OntoEnrich

OntoEnrich supports ontologists to analyze ontologies from the lexical perspective. Fig. 1 shows the components of the OntoEnrich platform. Each user has a profile with personal execution and storage constraints. The LA of an ontology starts by the lexical analysis, which produces a set of LRs. This is done once for each ontology and it is automatically performed. Its execution time depends on the ontology size. The results are stored in a reusable XML file. The algorithms and methods are encapsulated in the Onto-Enrich Java API, which uses external libraries like the OWL API (http://owlapi.sourceforge.net/) for manipulating ontologies, and the Stanford CoreNLP (http://nlp.stanford.edu/software/corenlp.shtml) for tokenization, part-of-speech tagging and lemmatization of labels, which is used for building a graph of tokens used for detecting the LRs. Further details about how the LRs are detected can be found in [3].

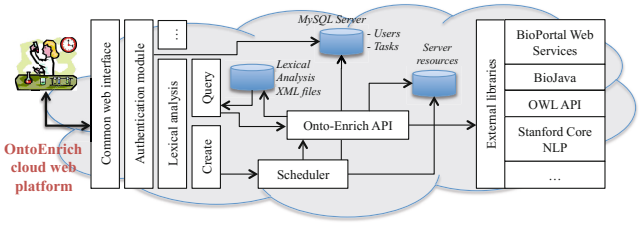

Fig. 1. General component architecture of the OntoEnrich platform

Each LR has some *general descriptors* like: its text, if it is a class of the ontology, the labels that exhibit it and the type of word/s that it is. For example, "binding" appears in 1222 labels of the GO-MF, it is a class and it is a noun or the nominal form of the verb "to bind" in the labels that exhibit it. If an LR has different forms (e.g., noun vs. verb) it can be split in two according to the role played. Other *advanced features* associated with each LR are:

- **Sub/super-regularities:** a sub-regularity is a sub-sequence of an LR, while a super-regularity is obtained by extending the LR in any direction. Their analysis helps to refine LRs. For example, "receptor binding" (339 repetitions) is more specific than "binding".
- **Generalization of regularities:** the alignment of the labels that exhibit an LR helps in the automatic identification of patterns like "translation X factor activity" several LRs. For this, we have adapted bioinformatics multiple sequence alignment algorithms.
- The **cross-product extension metric** informs about the degree of enrichment of an LR using matches obtained from other ontology selected by the user, and the **localization and modularization metrics** inform about the distribution in the asserted hierarchy of classes that exhibit an LR. For example, if an LR is a class in the ontology: (1) this class should be the common ancestor of the classes that exhibit such an LR, or (2) these classes should be linked with other type of relationship.

3 Example of the Inspection of Lexical Regularities

We illustrate the OntoEnrich platform with the GO-MF, and how the inspection of LRs might help to identify deviations or lexico-syntactic patterns like those manually detected in [2] (see Fig. 2). We omit the LA step. For further details, please check the tutorials in our website (see Fig. 2-1).

Fig. 2 shows the screenshots for the "binding" and "forming" LRs. Fig. 2-3 shows the information of the LR under inspection. We can navigate through the LRs (see Fig. 2-8). In Fig. 2-4 the *general descriptors* of the active LR are shown, and the labels that exhibit the LR can be explored in Fig 2-5. *More complex features* of the LR are analyzed independently and they are chosen using Fig. 2-6. Panel 5 shows the labels in which the LR appears. Panel 7 contains information about the super-patterns, sub-patterns, or alignment of labels, depending on the option selected in Panel 6.

Use Case 1 - "binding" (Fig. 2 left): this LR is quite general, so the inspection of the super-regularities can be useful. For example, there are 23 classes that exhibit the super-regularity "ion binding", which is a class in the ontology; however, the least common sub-summer of these 23 classes is "binding" instead of "ion binding", which suggests the inspection of the labels that exhibit "ion binding" for discarding that there are irregularitites in the naming of the labels. Hence, this analysis could serve to inspect the correlation between the lexical regularities and relationships between the corresponding classes.

Use Case 2 -"forming" (Fig. 2 right): this LR is recognized as a verb by the NLP modules and, according to [1], verbs usually codify semantic relationships. If we align and analyze the labels that exhibit this LR, the first 6 labels could be generalized as: 'ligase activity, forming ?x'. Then, if ?y represents classes that follow such a pattern, these classes can be enriched with the axioms *'?y subClassOf "ligase activity"* and *'?y subClassOf enables some (forming some ?x)'*,

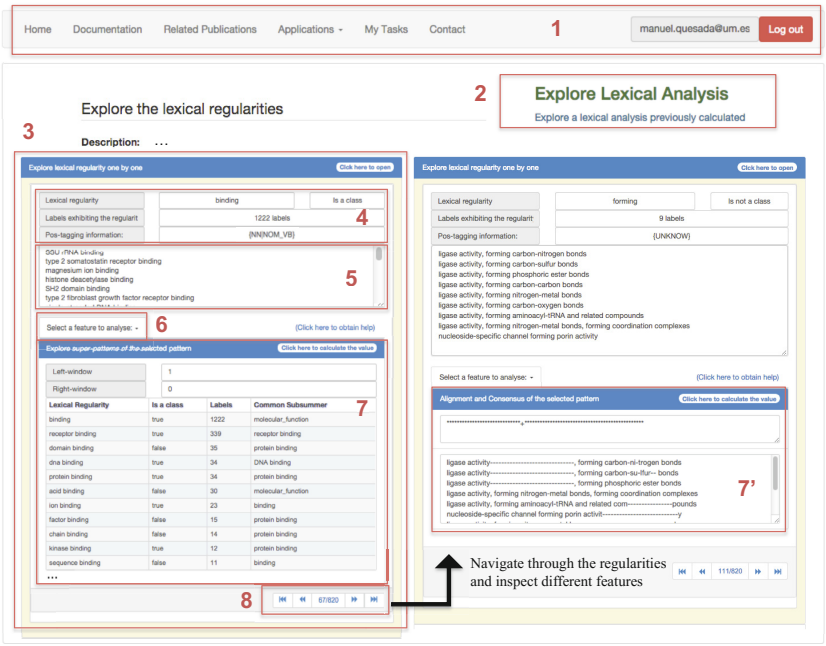

Fig. 2. Example of the online inspection of lexical regularities (http://sele.inf.um.es/ontoenrich/files/ekaw2014ontoenrichImg.pdf)

where the LR is created as an object property. However, the alignment of labels that exhibit the LR does not obtain consensus as "nucleoside-specific channel forming porin activity" does not follow the pattern "Y, forming X". In the other two labels several elements are formed, so two axioms with an AND clause might be created.

In general, this information might be used to debug the ontology in case abnormalities are found and to automatically generate Ontology Design Patterns (http://ontologydesignpatterns.org/), which can be implemented in OPPL scripts (http://oppl2.sourceforge.net/) to refine the class hierarchy.

4 Availability and Future Work

OntoEnrich is available at http://sele.inf.um.es/ontoenrich. We hope to extend it with algorithms that help in the automatic detection of lexico-syntactic patterns and its codification as OPPL scripts that create the ontology axioms.

Acknowledgments. This project has been possible thanks to the Spanish Ministry of Science and Innovation and the FEDER Programme through grant TIN2010-21388-C02-02 and fellowships BES-2011-046192 (MQM) and EEBB-I-14-08700 (MQM), and by the Fundación Séneca (15295/PI/10).

References

1. Aussenac-Gilles, N., Jacques, M.-P.: Designing and evaluating patterns for relation acquisition from texts with Caméléon. Terminology **14**(1), 45–73 (2008)
2. Fernandez-Breis, J.T., Iannone, L., Palmisano, I., Rector, A.L., Stevens, R.: Enriching the gene ontology via the dissection of labels using the ontology pre-processor language. In: Cimiano, P., Pinto, H.S. (eds.) EKAW 2010. LNCS, vol. 6317, pp. 59–73. Springer, Heidelberg (2010)
3. Quesada-Martínez, M., Fernández-Breis, J.T., Stevens, R.: Lexical characterization and analysis of the BioPortal ontologies. In: Peek, N., Marín Morales, R., Peleg, M. (eds.) AIME 2013. LNCS, vol. 7885, pp. 206–215. Springer, Heidelberg (2013)