Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 171-174 recibido 31-03-2017 revisado 11-05-2017 aceptado 29-05-2017

OntoEnrich: Una plataforma para el análisis léxico de ontologías orientado a su enriquecimiento axiomático

OntoEnrich: A platform for the lexical analysis of ontologies focused on their axiomatic enrichment

Manuel Quesada-Martínez
Universidad de Murcia / IMIB-Arrixaca, Facultad de Informática, CP 30100 Murcia
manuel.quesada@um.es

Dagoberto Castellanos-Nieves
Universidad de La Laguna, Departamento Ingeniería Informática y de Sistemas, CP 38271 La Laguna
dcastell@ull.es

Jesualdo T. Fernández-Breis
Universidad de Murcia / IMIB-Arrixaca, Facultad de Informática, CP 30100 Murcia
jfernand@um.es

Resumen: OntoEnrich es una plataforma online para la detección automática y análisis de regularidades léxicas encontradas en las etiquetas asociadas a los conceptos de una ontología. Un análisis guiado por estas regularidades permite explorar diferentes aspectos léxico/semánticos, como puede ser la aplicación de los principios del OBO Foundry en el caso de ontologías biomédicas. El objetivo de esta demostración es presentar casos de uso obtenidos al aplicar la herramienta en ontologías relevantes como Gene Ontology o SNOMED CT. Mostraremos cómo dicho análisis permite identificar semántica oculta a partir de contenido descrito en lenguaje natural (apto para humanos), y cómo podría ser usado para enriquecer la ontología creando nuevos axiomas lógicos (aptos para máquinas).
Palabras clave: Ontologías, PLN, enriquecimiento axiomático, análisis léxico

Abstract: We present OntoEnrich, an online platform for the automatic detection and guided analysis of lexical regularities in ontology labels. An analysis guided by these regularities permits users to explore different lexical and semantic aspects as the application of the OBO Foundry principles in biomedical ontologies. The goal of this demonstration is to show some use cases obtained after applying OntoEnrich in two relevant biomedical ontologies such as Gene Ontology and SNOMED CT. Thus, we will show how the performed analysis could be used to elucidate hidden semantics from the natural language fragments (human-friendly), and how this could be used to enrich the ontology by generating new logical axioms (machine-friendly).
Keywords: Ontologies, NLP, axiomatic enrichment, lexical analysis

1 Introducción

En los últimos años, el interés de la comunidad biomédica en el uso de ontologías ha motivado un crecimiento continuo en la cantidad de ontologías disponibles. Por ejemplo, el repositorio BioPortal[1] contenía más de 500 ontologías en Marzo de 2017, y cerca de 8 millones de clases. Brevemente, una ontología, entendida como artefacto software, está compuesta por clases, propiedades e instancias; y contiene axiomas lógicos que permiten inferir nuevo contenido mediante el uso de razonadores (Guarino, 1998). A menudo, las ontologías biomédicas son desarrolladas por equipos multidisciplinares de ingenieros de ontologías y expertos en el dominio. El lenguaje natural favorece la comunicación entre humanos, sin embargo, éste debería ser también expresado como axiomas lógicos para que sea interpretable por los razonadores.

Comunidades como el OBO Foundry propone principios de buenas prácticas para crear conjuntos de ontologías ortogonales (Smith et al., 2007). Por ejemplo, a cada concepto se le asocia una `label` que lo debe describir sin ambigüedad usando: lenguaje natural y un nombrado sistemático. Comprobar si se sigue el principio "*lexically suggest, logically define*" (Rector y Iannone, 2012) ofrecería información sobre la consistencia entre el contenido expresado en las `labels` y el modelo lógico definido por los axiomas. Por

[1] https://bioportal.bioontology.org

ISSN 1135-5948

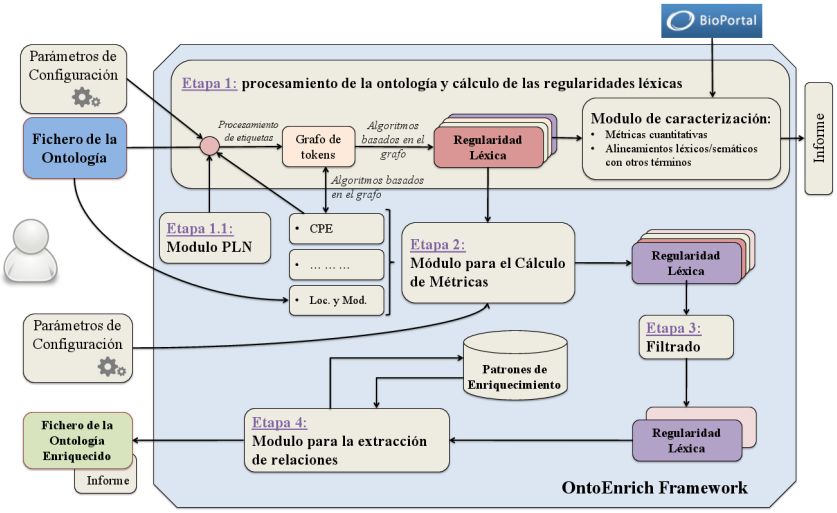

Figura 1: Descripción de la metodología aplicada por OntoEnrich

ejemplo, el nombrado de las clases `binding` y `receptor binding` sugiere la necesidad de una relación jerárquica entre ellas. Si la relación existe, el principio se estaría cumpliendo. En otro caso, la regularidad léxica 'binding' permitiría identificar semántica oculta (Third, 2012) aplicable en el enriquecimiento de la ontología con nuevos axiomas (Fernandez-Breis et al., 2010).

Tradicionalmente, el procesamiento del lenguaje natural ha sido aplicado al análisis de textos para crear o enriquecer ontologías (Brewster et al., 2009; Buitelaar, Cimiano y Magnini, 2005). Aquí nos centramos en analizar las `labels` que son a menudo descripciones muy breves. El enriquecimiento de ontologías basado en `labels` se ha abordado individualmente para ontologías específicas y aplicando patrones de enriquecimientos predefinidos; (Mungall et al., 2011) y (Golbreich, Grosjean y Darmoni, 2013) son algunos ejemplos. Nuestra hipótesis es que ayudar a los desarrolladores de ontologías en el análisis de regularidades léxicas podría contribuir a garantizar la calidad de las mismas mediante su enriquecimiento, y aumentar su utilidad al ser aplicadas en proyectos reales como (Aguilar et al., 2016).

2 *El framework OntoEnrich*

OntoEnrich implementa una metodología para el enriquecimiento de ontologías biomédicas basado en el análisis léxico de sus etiquetas (Quesada-Martínez, 2015). La Figura 1 muestra sus principales etapas y son brevemente comentadas a continuación. El método acepta una ontología como entrada. Durante la etapa 1, la ontología se procesa automáticamente para obtener las `labels`, y se aplica un proceso de tokenización y lematización usando la librería Stanford Core NLP[2] (etapa 1.1). También se obtienen las etiquetas gramaticales de cada uno de los tokens así como nominalizaciones de verbos utilizando los recursos ofrecidos por el SPECIALIST lexicon[3]. Toda esta información se almacena en un grafo, que nos permite hacer diferentes tipos de consultas. Una regularidad léxica (RL) es un conjunto de tokens consecutivos repetidos en diferentes etiquetas de la ontología. En esta primera etapa, cada RL se utiliza para calcularle un conjunto de métricas cuantitativas que permiten la caracterización léxica de la ontología produciendo un informe de salida. Algunas de estas métricas utilizan algoritmos de alineamiento que permiten identificar elementos ya definidos en la propia ontología o en otras externas; esto pretende promover la reutilización de conceptos entre la comunidad biomédica. Siguiendo con el ejemplo, la RL 'binding' aparece en 1222 labels de la ontología de funciones moleculares de Gene Ontology (GOMF), y es la etiqueta de una clase. En la etapa 2, se propone el uso de métricas avanzadas que relacionan las RLs con diferentes aspectos semánticos de la ontología. Por ejemplo, la métrica de productos cruzados informa sobre el grado de enriquecimiento de una regularidad léxica usando los alineamientos obtenidos por las clases que la

[2]http://nlp.stanford.edu/software/corenlp.shtml
[3]https://specialist.nlm.nih.gov/lexicon/

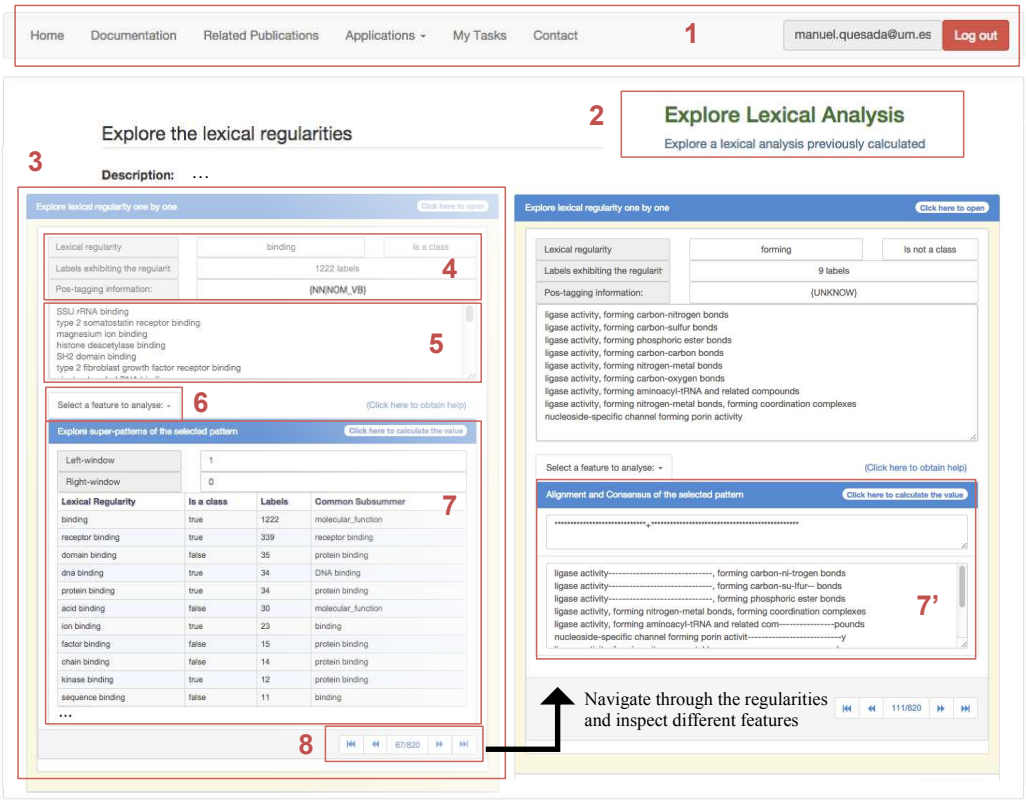

Figura 2: Ejemplo de la inspección de las regularidades léxicas "binding" y "forming"

exhiben. Otro ejemplo, las funciones de similitud semántica son aplicados para contextualizar aquellas clases que exhiben una RL teniendo en cuenta la jerarquía definida por las relaciones (métricas de localización y modularidad). Estas métricas pretenden cuantificar la respuesta a preguntas como ¿cuántas clases que exhiben 'binding' son descendientes o están relacionadas con él? El cálculo de las métricas puede requerir la configuración de un conjunto de parámetros de entrada por parte del usuario. Las métricas permiten definir filtros que reducen el conjunto de RLs a aquellas que cumplen ciertas propiedades (etapa 3).

También se puede utilizar etiquetado gramatical de los tokens como filtro. Por ejemplo, 'binding' es la nominalización del verbo "to bind" y esta información podría derivar en la generación del patrón de enriquecimiento "X binding", el cual añade a las clases que exhiben la RL el axioma *"subClassOf enables some (binds some ?x)"*. El patrón de enriquecimiento se define usando el lenguaje OPPL[4] y puede ser incluido en repositorios de patrones reutilizable de diseño de ontologías[5]. Esta transformación sería el último paso de la metodología (etapa 4). Como resultado de la ejecución de dichos patrones se obtendría la ontología enriquecida.

3 *Análisis léxicos a través de la plataforma online*

OntoEnrich está disponible como aplicación web y encapsulado en una librería Java integrable con otros programas[6]. El objetivo de la web es facilitar el análisis y la interacción para usuarios sin conocimientos técnicos. Un usuario debe registrarse. El tiempo dedicado al análisis léxico de una ontología dependerá de su tamaño y de los parámetros de entrada seleccionados. Por ello el usuario programa el análisis y una vez finalizado su cálculo se almacena en un fichero XML reutilizable.

[4]https://github.com/owlcs/OPPL2
[5]http://ontologydesignpatterns.org/
[6]http://sele.inf.um.es/ontoenrich

La Figura 2 muestra una captura de la aplicación. Usando el menú superior el usuario puede navegar sobre los distintos análisis disponibles. En este caso mostramos un extracto de la información relativa a las RLs "binding" y "forming" encontradas en GOMF. Su análisis interactivo permite identificar desviaciones o patrones de enriquecimiento como el comentado en la sección anterior. En esta demostración se pretende mostrar tres workflows que han sido diseñados para la aplicación de OntoEnrich a Gene Ontology (GO) y SNOMED CT. Los workflows proponen un conjunto de pasos usando métricas y filtros que permiten analizar:

- GO: si las RLs (alineadas con clases) deberían ser el ancestro común de todas las clases que las exhiben. Vídeotutorial[7].
- SNOMED CT: usar el etiquetado gramatical para detectar RLs que son adjetivos y que siguiendo el principio *lexically suggest, logically define* deberían estar relacionados con `qualifier values` definido en la ontología. Vídeotutorial[8].
- GO: formateo de las etiquetas que exhiben una regularidad para obtener expresiones regulares convertibles en patrones de enriquecimiento. Vídeotutorial[9].

Estos y otros ejemplos están disponibles en la sección de documentación de la página web de Ontoenrich.

4 Conclusiones

OntoEnrich es una plataforma integrada que permite el análisis de regularidades léxicas en ontologías. El uso de métricas permite al usuario centrarse en diferentes aspectos que pueden contribuir a garantizar la calidad de las ontologías identificando desviaciones o puntos de mejora basados en el contenido descrito en lenguaje natural.

Agradecimientos

Este trabajo ha sido posible gracias al Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional (FEDER), a través del proyecto TIN2014-53749-C2-2-R, y a la Fundación Séneca a través del proyecto 19371/PI/14.

[7] https://tinyurl.com/mhmnbhv
[8] https://tinyurl.com/kgpx9y8
[9] https://tinyurl.com/lcqfdl3

Bibliografía

Aguilar, C. A., O. Acosta, G. Sierra, S. Juárez y T. Infante. 2016. Extracción de contextos definitorios en el área de biomedicina. *Procesamiento del Lenguaje Natural*, 57:167–170.

Brewster, C., S. Jupp, J. Luciano, D. Shotton, R. D. Stevens y Z. Zhang. 2009. Issues in learning an ontology from text. *BMC bioinformatics*, 10(5):S1.

Buitelaar, P., P. Cimiano y B. Magnini. 2005. *Ontology learning from text: methods, evaluation and applications*, volumen 123. IOS press.

Fernandez-Breis, J., L. Iannone, I. Palmisano, A. Rector y R. Stevens. 2010. Enriching the Gene Ontology via the dissection of labels using the ontology pre-processor language. *Know. Engineering and Management by Masses*, páginas 59–73. Springer.

Golbreich, C., J. Grosjean y S. J. Darmoni. 2013. The Foundational Model of Anatomy in OWL 2 and its use. *Artificial Intelligence in Medicine*, 57(2):119–132.

Guarino, N. 1998. Formal Ontology in Information Systems: Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy, páginas 3-15. IOS Press.

Mungall, C. J., M. Bada, T. Z. Berardini, J. Deegan, A. Ireland, M. A. Harris, D. P. Hill y J. Lomax. 2011. Cross-product extensions of the Gene Ontology. *Journal of Biomedical Informatics*, 44(1):80–86.

Quesada-Martínez, M. 2015. *Methodology for the enrichment of biomedical knowledge resources*. Ph.D. tesis, Depto. de Informática y Sistemas. Univ. de Murcia.

Rector, A. y L. Iannone. 2012. Lexically suggest, logically define: Quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED. *Journal of Biomedical Informatics*, 45:199–209.

Smith, B., M. Ashburner, C. Rosse, J. Bard, W. Bug y others. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251-1255.

Third, A. 2012. "Hidden Semantics": What Can We Learn from the Names in an Ontology? En *Proceedings of the 7th International Natural Language Generation Conference*, INLG '12, páginas 67–75. ACL.