Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 139-142 recibido 14-03-2016 revisado 13-04-2016 aceptado 28-04-2016

DBpedia del gallego: recursos y aplicaciones en procesamiento del lenguaje

Galician DBpedia: resources and applications in language processing

Miguel Anxo Solla Portela
Universidade de Vigo
Grupo TALG
miguelsolla@uvigo.es

Xavier Gómez Guinovart
Universidade de Vigo
Grupo TALG
xgg@uvigo.es

Resumen: En esta presentación, describimos la metodología utilizada para la creación de la DBpedia del gallego y algunas de sus aplicaciones para el procesamiento lingüístico en los ámbitos del reconocimiento de entidades y de la extracción léxica.
Palabras clave: DBpedia, Wikipedia, WordNet, datos enlazados abiertos, web semántica

Abstract: In this presentation, we review the methodology used in the development of the Galician DBpedia and some of its applications for language processing in the fields of entity recognition and lexical extraction.
Keywords: DBpedia, Wikipedia, WordNet, linked open data, semantic web

1 Introducción

En este artículo[1] se describe la metodología seguida en la creación de la DBpedia del gallego y algunas de sus aplicaciones en el campo del procesamiento del lenguaje. La construcción de este recurso se realizó gracias a la financiación de la Red de Investigación *Tecnoloxías e análise dos datos lingüísticos*, orientada al desarrollo de recursos para el procesamiento lingüístico del gallego, siendo uno de sus objetivos principales la puesta en marcha de nuevas aplicaciones y herramientas con tecnologías de base semántica.

La DBpedia[2] (Lehmann et al., 2015) es un proyecto internacional para crear una versión estructurada de los contenidos de la Wikipedia[3] y publicarla libremente en Internet entrelazada con el conjunto de bases de conocimiento que constituyen la web semántica.

La DBpedia permite realizar consultas complejas a partir del conjunto de datos derivados de la Wikipedia y permite enlazar estos datos con otros conjuntos de datos que hay en la web, siguiendo las especificaciones para los datos enlazados abiertos (Linked Open Data)[4] establecidas por el W3C (World Wide Web Consortium) (Auer et al., 2007).

2 Recursos

La DBpedia del gallego, desarrollada y mantenida por el Grupo TALG (*Tecnoloxías e Aplicacións da Lingua Galega*) de la Universidade de Vigo, contiene 11 millones de tuplas semánticas extraídas a partir de toda la información contenida en la Galipedia[5] y está alojada en el subdominio oficial de dbpedia.org correspondiente a la lengua gallega[6].

La elaboración de la DBpedia del gallego supuso la adaptación de la aplicación de extracción de los datos procedentes de los ficheros *dump* de la Wikipedia, de Wikimedia Commons[7] y de Wikidata[8] para que funcionase satisfactoriamente con los datos procedentes de la Galipedia. Las modificaciones realizadas en el código de la aplicación se pueden consultar en Github[9] y han sido ya implementadas en la aplicación principal de extracción de la DBpedia[10].

[1] Esta investigación se realizó en el marco de la Red de Investigación *Tecnoloxías e análise dos datos lingüísticos* financiada por la Consellería de Cultura, Educación e Ordenación Universitaria de la Xunta de Galicia, ref. CN 2014/007.

[2] `http://dbpedia.org`

[3] `http://wikipedia.org`

[4] `https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

[5] `http://gl.wikipedia.org`

[6] `http://gl.dbpedia.org`

[7] `https://commons.wikimedia.org`

[8] `https://www.wikidata.org`

[9] `https://github.com/galician/extraction-framework/`

[10] `https://github.com/dbpedia/extraction-framework/`

Igualmente, con el mismo objetivo de creación del recurso, se elaboraron los ficheros de conversión (*mappings*) necesarios para obtener información estructurada a partir de las *infoboxes* y de las cajas de navegación de la Galipedia[11]. Aunque esta tarea se halla todavía en curso de finalización, la cobertura alcanzada con el trabajo ya realizado resulta bastante amplia, como se puede comprobar en las estadísticas disponibles de los *mappings* de la DBpedia[12]. El conjunto de datos se ha completado, además, con la extracción de los resúmenes de los artículos de la Galipedia ligados a cada recurso.

Los ficheros RDF de la DBpedia del gallego generados a partir de la Galipedia, pueden ser libremente descargados desde el sitio de la DBpedia[13], y sus contenidos pueden consultarse y visualizarse en la web del grupo mediante las aplicaciones Lodview[14] y LodLive[15] (ambas localizadas en gallego como parte del proyecto), utilizando la interfaz adaptada de la propia DBpedia[16] o a través del punto de acceso Virtuoso SPARQL a los datos estructurados[17].

La publicación del punto de acceso SPARQL propició también el modelado en formato de datos enlazados abiertos de Galnet[18] (Solla Portela y Gómez Guinovart, 2015), el WordNet 3.0 del gallego desarrollado por el Grupo TALG que forma parte de la distribución del Multilingual Central Repository (MCR) (González Agirre, Laparra, y Rigau, 2012). La consulta de la versión RDF de Galnet se encuentra disponible a través del servidor SPARQL de la DBpedia del gallego utilizando el grafo `http://sli.uvigo.gal/rdf_galnet`.

El diseño de la estructura de los datos RDF se basó en la versión 3.1 del WordNet de Princeton[19], siguiendo el modelo lemon[20], con ligeras modificaciones respecto al original para poder incorporar los enlaces con las clasificaciones semánticas y ontologías presentes en el MCR y Galnet[21] y mantener su naturaleza plurilingüe a través de un índice interlingüístico (ILI). Además, con el fin de ampliar su cobertura a consultas externas, se alineó cada synset con el correspondiente en la versión 3.1 de Princeton y con la versión 3.0 en formato lemonUby[22]. El resultado de este alineamiento conlleva la compatibilidad del índice interlingüístico de WordNet presente en el MCR con innumerables fuentes de datos enlazados que ya se encuentran disponibles en la web semántica.

3 Aplicaciones

3.1 DBpedia Spotlight

Una vez elaborados los recursos y habilitado el acceso abierto a los datos estructurados, se desarrolló una versión adaptada al gallego de la aplicación DBpedia Spotlight (Daiber et al., 2013) para poder ofrecer una primera herramienta de explotación inmediata de los datos de la DBpedia del gallego en el campo del procesamiento del lenguaje.

DBpedia Spotlight es una utilidad para la anotación de textos con referencias a los conceptos de la DBpedia. La identificación en contexto de las formas relativas a los conceptos se realiza mediante un sistema adaptable que localiza y desambigua de forma automática las menciones a recursos de la DBpedia presentes en el lenguaje natural. En este sentido, la identificación de entidades llevada a cabo por DBpedia Spotlight posee un alcance menos restringido que el reconocimiento de entidades nombradas, habitualmente limitado a ciertas categorías predefinidas como personas, organizaciones y lugares.

La adaptación al gallego de DBpedia Spotlight realizada en el marco de este proyecto identifica y anota en los textos las referencias a conceptos de la DBpedia del gallego, y puede utilizarse libremente desde su interfaz de usuario[23] o como servicio web[24].

[11] `http://mappings.dbpedia.org/index.php/Mapping_gl`

[12] `http://mappings.dbpedia.org/server/statistics/gl/`

[13] `http://downloads.dbpedia.org/2015-10/core-i18n/gl/`

[14] `http://sli.uvigo.gal/dbpedia/lodview/`

[15] `http://sli.uvigo.gal/dbpedia/lodlive/`

[16] `https://github.com/dbpedia/dbpedia-vad-i18n`

[17] `http://gl.dbpedia.org/sparql/`

[18] `http://sli.uvigo.gal/galnet/`

[19] `http://wordnet-rdf.princeton.edu`

[20] `http://lemon-model.net`

[21] Concretamente, los WordNet Domains (Bentivogli et al., 2004), la ontología Adimen-SUMO (Álvez, Lucio, y Rigau, 2012), la Top Ontology (Álvez et al., 2008), los Basic Level Concepts (Izquierdo, Suárez, y Rigau, 2007) y los epinónimos (Solla Portela y Gómez Guinovart, 2015)

[22] `http://lemon-model.net/lexica/uby/wn/`

[23] `http://sli.uvigo.gal/dbpedia/spotlight/`

[24] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Web-service`

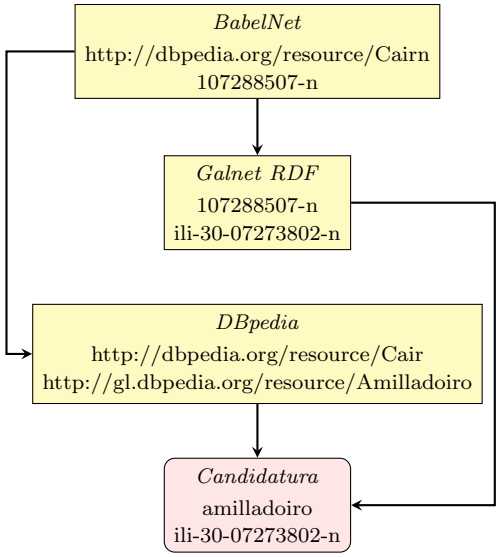

Figura 1: Extracción de variantes (1).

3.2 Extracción léxica

Para poder comprobar las posiblidades de explotación de estos recursos LOD en otras tareas de procesamiento del lenguaje, diseñamos dos experimentos de extracción léxica basados en la DBpedia dirigidos a la ampliación del WordNet del gallego. En el primer experimento de extracción, a parte de la DBpedia del gallego y de Galnet, usamos como fuente LOD remota la versión RDF de BabelNet[25]. El objetivo del experimento consiste en aumentar la cobertura de Galnet mediante variantes gallegas procedentes de la DBpedia limitándose a los synsets de Galnet que aún non tuvieran variantes gallegas.

En primer lugar, se obtuvieron de BabelNet los identificadores de sentido de WordNet 3.1 ligados a recursos de la DBpedia en inglés. El número de alineamientos identificador–recurso obtenidos mediante esta fuente ascendió a 7.796. Segundo, se obtuvieron de Galnet los ILIs de WordNet 3.0 correspondientes a los identificadores de sentido de WordNet 3.1 procedentes de BabelNet. Simultáneamente, se obtuvieron de la DBpedia del gallego los recursos gallegos correspondientes a los recursos de la DBpedia del inglés procedentes de BabelNet[26]. Por último, se identifican los synsets de Galnet correspondientes a los ILIs de WordNet 3.0 obtenidos y se proponen como candidatos a variante los recursos relacionados de la DBpedia del gallego.

Con esta estrategia se consiguieron 910 candidaturas con variantes nominales que apuntaban a synsets que todavía no tenían ninguna variante en gallego. El índice de precisión obtenido en el experimento de extracción, tras su revisión humana, alcanzó el 82,3 %, como se refleja en los resultados de la Tabla 1. Durante la revisión se observó además que, salvo en algunos casos aislados en los que la equivalencia entre idiomas en la DBpedia no es correcta, en la mayor parte de los casos en los que no se puede establecer la validez, el origen del error se encuentra en la inadecuación del alineamiento entre el recurso de la DBpedia y el identificador de WordNet 3.1 en BabelNet. La Figura 1 ilustra este proceso de extracción de variantes de Galnet a partir de los recursos LOD de la DBpedia, BabelNet y Galnet con un ejemplo de candidatura aceptada[27].

Variantes evaluadas	910	
Aceptadas	749	82,3 %
Rechazadas	161	17,7 %

Tabla 1: Evaluación de las candidaturas (1).

En un segundo experimento, exploramos la adquisición de variantes a partir de las equivalencias interlingüísticas de la DBpedia y de las variantes interlingüísticas presentes en los synsets del MCR. Partiendo de los synsets sin variante en gallego, se compararon las variantes existentes en catalán, euskera, portugués, español e inglés con los recursos de la DBpedia para cada una de estas lenguas, a fin de proponer candidaturas de nuevas variantes para el gallego (Figura 2). Con este método se generaron 2.194 candidaturas a partir de recursos con al menos una variante coincidente en alguna de las lenguas de los wordnets del MCR, con un índice de precisión tras la revisión humana del 88,3 % (Tabla 2).

Variantes evaluadas	2.194	
Aceptadas	1.937	88,3 %
Rechazadas	257	11,7 %

Tabla 2: Evaluación de las candidaturas (2).

[25] `http://babelnet.org/rdf/`

[26] Es preciso tener en cuenta que las tuplas de equivalencias interlingüísticas de la DBpedia se generan con el mismo código de extracción de información estructurada que se utiliza para la Wikipedia, pero se toman como fuente los datos procedentes de Wikidata.

[27] `http://sli.uvigo.gal/galnet/galnet_var.php?version=dev&ili=ili-30-07273802-n`

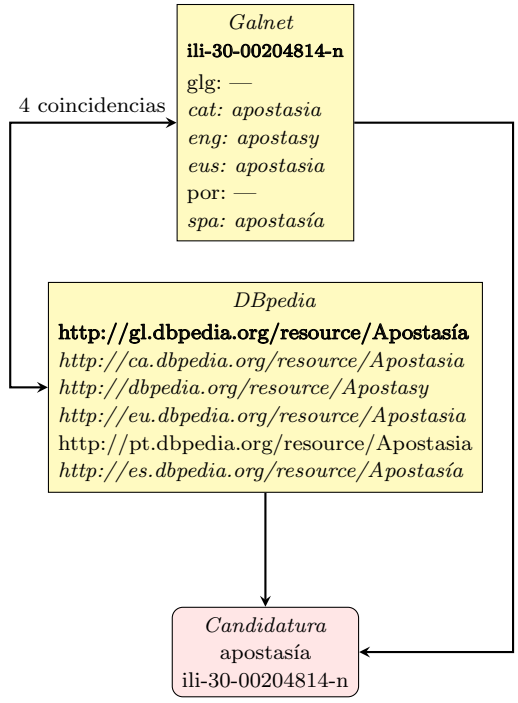

Figura 2: Extracción de variantes (2).

Las variantes aceptadas en estos dos experimentos fueron incorporadas al WordNet del gallego y pueden ser consultadas a través de su interfaz seleccionando como experimento *dbpedia*[28]. Ambas estrategias de extracción léxica pueden ser aplicadas, utilizando los mismos recursos, para sugerir candidaturas de variantes en cualquiera de las lenguas incluidas en los wordnets del MCR.

4 Conclusiones

La publicación de la DBpedia del gallego representa un avance importante para la presencia de información estructurada en lengua gallega en la web semántica. El punto de acceso SPARQL garantiza su aprovechamiento público en aplicaciones derivadas, además de permitir su interacción con los recursos disponibles en otros servidores con tecnologías semánticas. La explotación de la base de conocimientos de la DBpedia del gallego, en combinación con otros recursos en la web semántica, permitirá sin duda dinamizar proyectos, diseñar investigaciones y generar aplicaciones de gran interés en el ámbito del procesamiento del lenguaje.

[28] `http://sli.uvigo.gal/galnet_rev/galnet.php?version=dev&experiment=dbpedia`

Bibliografía

Álvez, J., J. Atserias, J. Carrera, S. Climent, A. Oliver, y G. Rigau. 2008. Consistent annotation of EuroWordNet with the Top Concept Ontology. En *Proceedings of the 4th Global WordNet Conference*, Szeged. GWN.

Álvez, J., P. Lucio, y G. Rigau. 2012. Adimen-SUMO: Reengineering an Ontology for First-Order Reasoning. *International Journal on Semantic Web and Information Systems*, 8(4):80–116.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, y Z. Ives. 2007. Dbpedia: A nucleus for a web of open data. En *In 6th Int'l Semantic Web Conference, Busan, Korea*, págs. 11–15. Springer.

Bentivogli, L., P. Forner, B. Magnini, y E. Pianta. 2004. Revising WordNet domains hierarchy: Semantics, coverage, and balancing. En *Proceedings of COLING Workshop on Multilingual Linguistic Resources*, págs. 101–108, Geneva. ACL.

Daiber, J., M. Jakob, C. Hokamp, y P. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. En *Proc. of the 9th International Conference on Semantic Systems*.

González Agirre, A., E. Laparra, y G. Rigau. 2012. Multilingual Central Repository version 3.0. En *6th Global WordNet Conference*.

Izquierdo, R., A. Suárez, y G. Rigau. 2007. Exploring the Automatic Selection of Basic Level Concepts. En *Proc. of the International Conference on Recent Advances on Natural Language Processing*, págs. 298–302, Shoumen.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, y C. Bizer. 2015. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195.

Solla Portela, M. A. y X. Gómez Guinovart. 2015. Galnet: o WordNet do galego. Aplicacións lexicolóxicas e terminolóxicas. *Revista Galega de Filoloxía*, 16:169–201.