INVOCA: Consultando la Linked Open Data en Lenguage Natural *

INVOCA: Querying the Linked Open Data in Natural Language

Eduardo Lupiani, Victoriano Navarro
Juana M. Ruiz-Martínez
Rafael Valencia-García
Universidad de Murcia
elupiani@um.es, vna12763@um.es
jmruymar@um.es, valencia@um.es

Pedro J. Vivancos-Vicente
Juan S. Castejón-Garrido
VÓCALI Sistemas Inteligentes
pedro.vivancos@vocali.net
juans.castejon@vocali.net

Resumen: La finalidad del proyecto INVOCA, es el desarrollo de una interfaz de consulta en lenguaje natural para acceder a las distintas bases de conocimiento disponibles actualmente en la Web mediante la "Linked Open Data". Este proyecto ha sido desarrollado conjuntamente por la empresa VOCALI y el grupo TECNOMOD de la Universidad de Murcia.
Palabras clave: Sistema Pregunta-Respuesta, Procesamiento del Lenguaje Natural, Linked Open Data

Abstract: The main objective of the INVOCA project is the development of a natural language interface to query the knowledge bases available on the Internet in the Linked Open Data. This project has been developed between the VOCALI enterprise and the TECNOMOD research group of the University of Murcia
Keywords: Question-Answering, Natural Language Processing, Linked Open Data

1. Introducción y objetivos del proyecto

La *Linked Open Data*, promovida por el W3C, permite compartir datos estructurados en la Web. Se basa en la idea de que el valor y la utilidad de los datos se incrementa cuanto mayor sea el número de enlaces haya entre ellos. De este modo, la *Linked Data* utiliza la Web para crear determinados tipos de enlaces entre datos de diferentes fuentes. Para representar los datos estructurados de la Web, la *Linked Data* se basa en el modelo de datos RDF y, por otro lado, para relacionar las bases de conocimiento (BC) heterogéneas disponibles actualmente en Internet se basa en el uso de enlaces RDF. Algunas de las comunidades más populares que están siguiendo estas convenciones que permiten enlazar información estructurada, son la DBpedia, DBLP, GeoNames, YAGO y WordNet. El resultado es que cada BC almacenada en cada una de estas comunidades complementa y completa a las demás.

A la hora de consultar y recuperar información de estas BC, el principal problema que surge es que los usuarios deben conocer: (1) la sintáxis de RDF, (2) algún lenguaje formal para consultar estas BC (SPARQL, RDQL o RQL), y (3) la estructura y el vocabulario de BC de destino (FOAF, DBpedia, GeoNames, etc.). Estos lenguajes no son tan intuitivos como el lenguaje natural y requieren conocimientos previos por parte del usuario.

El principal objetivo de este proyecto es proporcionar una interfaz de consulta en lenguaje natural que permita el acceso a las distintas BC que conforman *Linked Open Data* al grueso de los usuarios.

2. Estado actual del proyecto

Hasta el momento se ha definido una arquitectura para consultar en lenguaje natural las BC que conforman la *Linked Open Data*, además se ha construido un prototipo centrado en la consulta de la DBpedia, junto con otras ontologías propias del grupo de investigación TECNOMOD, así como bases de datos relacionales que han sido mapeadas a RDF mediante D2RQ.

* Este proyecto ha sido financiado por el Instituto de Fomento de la Región de Murcia ref:2008.03.ID+I.0034

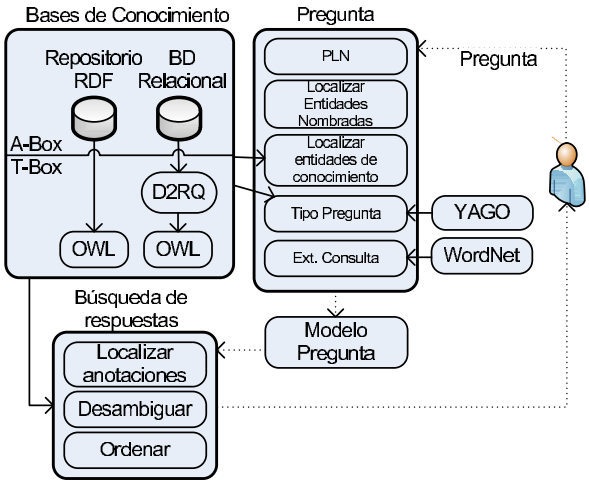

Figura 1: Arquitectura de INVOCA

A continuación se describe brevemente la arquitectura del sistema actual.

2.1. Arquitectura del Sistema INVOCA

El sistema INVOCA (ver Fig. **??**) está formado por tres módulos principales: Módulo de Bases de Conocimiento, el Módulo de procesamiento de la pregunta y el Módulo de búsqueda de respuesta. En primer lugar el usuario introduce una consulta en lenguaje natural, a continuación el módulo de procesamiento de la pregunta obtiene una representación sintáctico-semántica de la misma. El módulo de búsqueda de respuesta gestiona las posibles ambigüedades, accede a las BC y extrae la información que considera relevante para la consulta del usuario. Finalmente el sistema devuelve una respuesta. A continuación se describe brevemente cada uno de estos módulos.

2.2. Bases de Conocimiento

Este módulo se encarga de gestionar y proporcionar una interfaz común para las distintas BC que va a consultar el sistema. Cada BC está formada por una capa T-Box y una A-Box. La capa T-Box describe la estructura de clases, relaciones y restricciones de la BC, mientras que la capa A-Box está formada por las instancias de la BC. Este módulo permite actualmente la gestión de repositorios RDF SESAME, así como de ficheros OWL y bases de datos relacionales mapeadas a RDF mediante D2RQ. En el caso concreto de la DBpedia, esta BC ha sido integrada en un repositorio SESAME debido a la latencia del acceso online que presenta.

2.3. Procesamiento de la Pregunta

Este módulo procesa la pregunta en lenguaje natural y obtiene un modelo de la misma. El modelo de la pregunta, que se representa mediante OWL, ha sido definido específicamente para el proyecto. Dicho modelo contiene información sintáctica de la oración, el tipo de pregunta, los focos de la pregunta, las entidades nombradas y posibles entidades de conocimiento (URIs a recursos de las BC) junto con otra información.

Este módulo utiliza las herramientas GATE y Freeling, además de las ontologías YAGO y WordNet, para detectar los distintos elementos que describen sintáctico-semánticamente la pregunta.

2.4. Búsqueda de respuestas

Partiendo de un modelo de la pregunta este módulo, en primer lugar, resolverá las posibles ambigüedades que contenga dicho modelo en base a la información almacenada en las BC. A continuación se generan una o varias consultas SPARQL que permiten obtener la información que responde a la consulta del usuario. Una vez obtenido el resultado, éste se ordena en función a un score que indica la similitud semántica de la respuesta obtenida con respecto al modelo de la pregunta. Finalmente la respuesta se muestra al usuario.

3. Trabajo Futuro

Actualmente el sistema permite consultar las BC de DBpedia, así como otras ontologías descritas en OWL y en RDF en el repositorio. El siguiente paso será la mejora en la presentación de los resultados de las consultas que se muestran actualmente en el prototipo y la realización de una evaluación y validación exhaustiva de nuestro sistema en base a consultas que realicen usuarios sobre información de la DBpedia.

Se pretende también extender la búsqueda de respuestas incrementalmente a otras BC y ontologías como FOAF, SIOC o DBLP, así como añadir nuevas funcionalidades que permitan el multilingüismo en la consulta y recuperación de información.