Una arquitectura de integración de recursos léxicos de naturaleza heterogénea. Una aportación desde la perspectiva de la integración de datos.

Aitor Soroa Etxabe
Universidad del País Vasco / Euskal Herriko Unibertsitatea
649 P.K.
a.soroa@si.ehu.es

Resumen: En esta tesis se define una arquitectura para una federación de recursos léxicos de naturaleza heterogénea. Abordamos el problema de cunsultar diversas fuentes léxicas heterogéneas desde el punto de vista de la integración de datos, utilizando un lenguaje de consulta unificado.
Palabras clave: Integración de datos, bases de datos léxicas, integración de información léxica

Abstract: In this work we define an architecture for a federation of highly heterogeneous lexical information sources. We address the problem of querying very different existing sources of lexical information from the point of view of information integration, using for this purpose a unique and common query language.
Keywords: Data integration, lexical databases, lexical integration

1. Introducción

El área de los recursos léxicos[1] tiene cada vez más relevancia en el tratamiento automático del lenguaje natural (TLN), y ya nadie pone en duda que sin estos recursos es virtualmente imposible construir aplicaciones de TLN que trabajen sobre textos reales. Así, aparecen distintos proyectos que estudian la construcción de recursos léxicos altamente estructurados que contengan una gran riqueza de información léxica. En este contexto, aparecen dos términos relevantes: la reutilización de información léxica y la estandarización de la información léxica.

Se entiende por reutilización de información léxica al hecho de reutilizar recursos léxicos existentes a la hora de afrontar la construcción de un recurso nuevo. No obstante, la reutilización de la información léxica ha de hacer frente a diversos problemas, ya que los recursos léxicos suelen construirse para cumplir unos objetivos específicos, y muchas veces obedecen a teorías lingüísticas diferentes, e, incluso, contradictorias. Así, acceder a un recurso léxico requiere de un conocimiento profundo de la forma con la que el recurso estructura la información, así como del formato propio utilizado para almacenar los datos. Por lo tanto, un usuario que desee acceder a la información léxica en más de un recurso léxico ha de saber (Cunningham et al., 2000)

1. La organización de cada recurso para almacenar la información.
2. El lenguaje de consulta propio de cada sistema.

Los diferentes iniciativas para definir un estándar léxico que cubra todas las necesidades de representación léxica y a la vez almacenen la información de manera eficiente no han sido, en general, aceptados en la comunidad científica (Zajac, 1999). Como muestra de ello podemos citar el caso de WordNet, un recurso que no nació con la intención de ser un estándar léxico pero que ha sido adoptado como tal por muchos investigadores.

En esta tesis se propone construir un sistema de integración de recursos léxicos de naturaleza heterogénea. En este trabajo no se propone construir un gran repositorio centralizado en el que la información de las fuentes se duplique, tras un proceso de integración. En vez de eso, lo que se propone es que los datos residan en las fuentes y la integración se realice en el momento que el usuario consulte al sistema.

La autonomía de los recursos integrados es por tanto muy amplia, es decir, los recursos integrados no deben cambiar su diseño en ningún sentido para tomar parte en nuestro sistema. Así, decimos que las fuentes léxi-

[1] En este trabajo llamaremos recurso léxico.[a] cualquier repositorio de información léxica: Diccionarios, glosarios, bases de datos léxicas, etc.

cas integradas tienen "vida propia". De hecho, las fuentes léxicas pueden no "saber" que están siendo integradas.

2. Modelo Conceptual General

Nuestra propuesta de integración se basa en un modelo general de representación de información léxica, llamado Modelo Conceptual Global (MCG). El MCG, que está organizado como una jerarquía de clases y relaciones, y para representarlo hemos adoptado un sistema de representación basado en lógicas de descripciones, llamdo *NeoClassic*. El sistema permite hacer inferencias sobre la información almacenada, inferencias que serán necesarias para poder llevar a cabo la traducción de preguntas.

El MGC cumple dos funciones principales en el sistema. Por un lado, garantiza la comunicación con el usuario; así, el usuario utilizará los conceptos y relaciones del MCG para representar sus consultas. Por otro lado, el MCG será el esquema general sobre el cual se integraran las distintas fuentes léxicas del sistema: Cada esquema de los recursos integrados se relacionará con el MCG a través de reglas semánticas de integración.

Sobre cada recurso léxico hemos desarrollado un *wrapper*, que será el encargado de resolver la llamada heterogeneidad estructural, que se produce por los diferentes formatos y lenguajes de consulta empleados por los recursos para almacenar la información. El uso de *wrappers* nos permite hacer una abstracción y suponer que los recursos están compuestos por un conjunto de jerarquía de clases y relaciones, llamados Modelo Conceptual de los Recursos (MCR).

Para resolver la llamada heterogeneidad semántica —que se da porque cada recurso organiza y modela de forma diferente la información representada—, se han definido unas reglas semánticas, que relacionan cada recurso léxico integrado con el MGC. Para describir las reglas semánticas hemos adoptado el paradigma conocido en el área de la integración de información como "Lo local como vista" (Ullman, 1997), en la que cada clase y relación de cada recurso léxico se define en términos del MCG[2]. Esta estrategia nos permite que cada recurso se defina de forma independiente del resto, y, por lo tanto, garantiza la extensibilidad de todo el sistema: Al integrar un nuevo recurso no hay que modificar las reglas semánticas anteriormente definidas.

[2] Al contrario del llamado "Lo global como vista", en el que el esquema global (MGC) se define en términos de los modelos de las fuentes integradas.

3. ELHISA

En esta tesis se ha desarrollado el sistema ELHISA (Ezagutza Lexikal Heterogeneoen Integrazio SistemA, Sistema de Integración de Conocimiento Léxico Heterogéneo). Por medio del sistema ELHISA, el usuario —ya sea un usuario humano o una aplicación de TLN— conseguirá acceder a recursos léxicos diferentes de una manera unificada. Para ello, mandará una consulta al sistema, especificando qué información quiere obtener, en términos de clases y relaciones del MCG. El sistema, entonces, analiza la consulta del usuario, determina qué recursos son relevantes para responderla, y traduce la pregunta original al modelo conceptual propio de cada recurso. Después, envía las consultas traducidas a las fuentes léxicas pertinentes, recibe las respuestas de éstas, y presenta finalmente una respuesta unificada al usuario. En este prototipo se han integrado 5 recursos de naturaleza diferente: Un diccionario para uso humano (Euskal Hiztegia), dos bases de datos léxicas (EDBL y EDR) y dos bases de conocimiento léxicas (EuroWordNet e HIZTSUA).

Bibliografía

Cunningham, H., K. Bontcheva, W. Peters, y Y. Wilks. 2000. Uniform language resource access and distribution in the context of a General Architecture for Text Engineering (GATE). En *Proceedings of the Workshop on Ontologies and Language Resources (OntoLex'2000)*, Sozopol, Bulgaria, September.

Ullman, Jeffrey D. 1997. Information integration using logical views. En Afrati y Kolaitis, editores, *Database Theory—ICDT'97, 6th International Conference*, volumen 1186 de *Lecture Notes in Computer Science*, páginas 19–40, Delphi, Greece, 8–10 Enero. Springer.

Zajac, Rémi. 1999. On some aspects of lexical standardization. En *ACL/SIGLEX99 - Standardizing Lexical Resources*, University of Maryland, June 21,22.