Procesamiento del Lenguaje Natural, num. 29 (2002), pp.297-299 recibido 16-07-2002; aceptado 19-07-2002

LIQUID - Language Independent Querying for Information Discovery (IST-2000-25324)

Antonio S. Valderrábanos*
antonio.valderrabanos@madrid.sema.slb.com
Alexander L. Belskis*
alexander.belskis@madrid.sema.slb.com
Luis Iraola Moreno*
luis.iraola@ madrid.sema.slb.com
Jose Esteban Lauzán*
jfernando.esteban@madrid.sema.slb.com

*SchlumbergerSema sae
Calle Albarracín, 25
28037 Madrid - Spain

1. Objetivos

En el momento de comienzo de LIQUID, los sistemas de recuperación de información disponibles en el mercado no incluían soluciones para el problema del multilingüismo[1]. Así, el idioma en que se realiza la búsqueda determina el idioma de los documentos recuperados.

El objetivo de LIQUID es desarrollar una solución que proporcione acceso multilingüe para bases de datos textuales en dominios especializados del ámbito técnico y científico. LIQUID permite realizar una búsqueda en la lengua materna del usuario y obtener documentos relevantes en cualquier otra lengua disponible. El dominio elegido para desarrollar el primer prototipo es gastroenterología. Los idiomas considerados son: español, alemán, inglés y francés.

Además de desarrollar una solución específica para gastroenterología, LIQUID se propone desarrollar una metodología para proporcionar acceso multilingüe a bases de datos especializadas, de forma que esta solución sea traspasable a otros dominios e idiomas. En cuanto a las especificaciones, esta solución ha de cumplir los siguientes requisitos:

- ser trasladable a otros dominios, inicialmente dentro del campo de la medicina
- permitir la fácil incorporación de nuevas lenguas, particularmente del ámbito europeo
- utilizar recursos lingüísticos preexistentes, o desarrollables en un tiempo razonable, de forma que el desarrollo del sistema (así como la incorporación de nuevas lenguas o su desarrollo en otro entorno) sea viable
- funcionar como complemento, no como sustituto, de los sistemas monolingües de recuperación de información existentes

Este conjunto de especificaciones ha hecho que LIQUID se haya basado, siempre que ha sido posible, en recursos ya existentes o fácilmente desarrollables.

2. Resultados

LIQUID posee cuatro componentes fundamentales que describimos a continuación.

- una base de datos textual multilingüe
- un conjunto de términos multilingüe
- una red semántica independiente del idioma
- un sistema de extracción de terminología multilingüe (TExtractor)

La base de datos textual multilingüe. Este corpus contiene textos en español, alemán,

[1] Recientemente, empresas como Convera y Manning & Napier Information Services han comenzado a comercializar productos en el campo de la recuperación de información multilingüe.

ISSN 1135-5948

inglés y francés, en el ámbito de la gastroenterología. Está integrado tanto por textos paralelos (con versiones en varias lenguas) como no paralelos (con versiones en una sola lengua). La presencia de textos no paralelos es necesaria para verificar la capacidad de acceso multilingüe: cualquier documento que esté disponible en una sola lengua, típicamente inglés, debe ser accesible mediante consultas en cualquier otra lengua. Este corpus está integrado por fuentes como:

- Informes clínicos procedentes del proyecto ELCANO[2] (1.130 para inglés y español)
- Publicaciones científicas, y resúmenes de las mismas, como el *World Journal of Surgery*, obtenidas a través de PubMed (13.244 textos, en su mayoría en inglés)

El conjunto de términos multilingüe. Este glosario es paralelo, es decir, para cada concepto existe un término en todas las lenguas consideradas. Esta condición es imprescindible para posibilitar la correcta traducción de cualquier consulta de usuario a todos los idiomas. Este glosario está construido a partir de *The International Wordbook of Gastroenterology*, (R. Pounder y M. Hudson 1994), publicado por Radcliffe Medical Press. Contiene más de 5.000 términos por idioma, todos ellos con su respectiva traducción.

La red semántica. Esta red tiene la función de describir y establecer relaciones entre los conceptos relevantes del campo de la gastroenterología. En ella, cada nodo representa un concepto que se hace explícito mediante uno o más términos en cada idioma. Ejemplos de conceptos, ver Anexo - Tabla 1. Ejemplos de relaciones: ver Tabla 2. Ejemplos de tipos de conceptos: ver Tabla 3.

Estos tres componentes, (textos, glosario y red) se relacionan de la siguiente manera. El glosario representa el conjunto de términos (el conocimiento) contenido en los textos. Además, cada término del glosario están enlazado a uno o más nodos de la red semántica. De esta forma, a partir de un término, una consulta, es posible acceder a sus equivalentes en otros idiomas y a los documentos que lo contienen.

TExtractor. Dado que asegurar que un glosario es representativo de un conjunto de textos es siempre una tarea compleja, LIQUID incluye una herramienta de extracción de terminología, TExtractor, cuyo propósito es precisamente ampliar la cobertura del conjunto de términos inicial de forma que sea representativo del corpus. Asegurar la cobertura es crítico porque el conjunto de términos resultante se utiliza para indexar los textos. Para ello, TExtractor funciona de la siguiente manera:

- toma como punto de partida el conjunto de términos inicial y, a partir de éstos, genera nuevos términos, variantes de los anteriores;
- a continuación, estos nuevos términos son validados usando como referencia el corpus textual

Así, TExtractor incrementará la cobertura del conjunto de términos inicial, de forma que éste sea representativo de la terminología contenida en los textos. De acuerdo con la última evaluación realizada, TExtractor incrementa la cobertura de nuestro glosario en más de un 20% (exceptuando el alemán que se encuentra en un estadio anterior de desarrollo)[3].

Resultados: ver Tabla 4

3. Artículos y demostraciones presentados sobre LIQUID.

[2] ELCANO , European and Latin-American Countries Associated in a Networked database of Outstanding Guidelines in unusual clinical cases. INCO/DC, DG XIII.

[3] Actualmente, LIQUID se encuentra en su último semestre de sus dos años de duración. Existe ya un prototipo del sistema final y una versión final de TExtractor.

TExtractor: a multilingual terminology extraction tool, Sánchez Valderrábanos, A.; A. Belskis; L. Iraola Moreno. In Proceedings of the Human Language Technology Conference (HLT2002). San Diego, California. March 24-27, 2002. Pp. 379-384.

Multilingual Terminology Extraction and Validation, Sánchez Valderrábanos, A.; A. Belskis; L. Iraola Moreno. In Proceedings of the 3rd International Conference On Language Resources And Evaluation (LREC 2002). Las Palmas, Spain. 29-31 May, 2002. Pp. 2163-2170.

Anexo

Tabla 1 - Términos

Índice	Inglés	Alemán	Francés	Español
102	acid-producing	säureproduzierend	producteur d'aci	generador de ácido
104	acid-suppressive	Säure-unterdrückend	acide-suppressi	ácido-represivo

Tabla 2 - Relaciones

Name	Definition	Inverse link
IS_A	The basic hierarchical link in the Network. If one item "is a" another item then the first item is more specific in meaning than the second item.	INVERSE_IS_A
TREATS	Functional relation that applies a remedy with the object of effecting a cure or managing a condition.	TREATED_BY
COMPLICATES	Functional relation that causes more severity or complexity or results in adverse effects.	COMPLICATED_BY
PREVENTS	Functional relation that stops, hinders or eliminates an action or condition.	PREVENTED_BY

Tabla 3 - Tipos de conceptos para el nodo “Treatment”

Treatment	Drug, Surgical Procedure, Endoscopic Procedure, Radiological Procedure, Surveillance.

Tabla 4 - Resultados

Idioma	Términos iniciales	Términos generados y válidos	Incremento
Inglés	4253	971	22,83%
Español	4306	1053	24,45%
Francés	4310	1034	23,99%
Alemán	5243	201	3,83%