Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 347-348 recibido 03-05-2011 aceptado 22-05-2011

MULTIMEDICA: Extracción de información multilingüe en Sanidad y su aplicación a documentación divulgativa y científica[1]

MULTIMEDICA: Multilingual Information Extraction in Health domain and application to scientific and informative documents

Paloma Martínez
Departamento de Informática
Universidad Carlos III de Madrid
pmf@inf.uc3m.es

José C. González-Cristóbal
Dpto. Ing. de Sistemas Telemáticos
E.T.S.I. Telecomunicación
Universidad Politécnica de Madrid
josecarlos.gonzalez@upm.es

Antonio Moreno Sandoval
Departamento de Lingüística General,
Universidad Autónoma de Madrid
antonio.msandoval@uam.es

Resumen: El proyecto tiene como objetivo la definición y desarrollo de técnicas de extracción y búsqueda de información a partir de textos en el dominio biomédico, en particular, en dos líneas especiales: en primer lugar, el tratamiento de documentación científica en inglés sobre farmacología y en segundo lugar, el procesamiento de textos divulgativos sobre salud en idiomas como español y árabe. Estas técnicas de extracción incluyen el reconocimiento de entidades propias del dominio, aplicación de patrones y aprendizaje automático a la extracción de relaciones semánticas de interés y la integración de recursos léxicos específicos de sanidad (UMLS, SNOMED, etc.) para la mejora de aplicaciones. Por otro lado, la información extraída debe organizarse para su utilización en las herramientas de búsqueda e integración de información.
Palabras clave: extracción de información, tecnologías del lenguaje humano, recursos terminológicos

Abstract: The aim of this project is to define and develop information extraction and retrieval techniques based on texts from the medical domain. This will be carried out following two basic tasks: firstly, processing scientific documents in English about pharmacology, and secondly, processing informative texts about health topics in other languages such as Spanish and Arabic. These information extraction techniques include domain entities recognition, pattern recognition, machine learning for extracting semantic relations, and the integration of lexical resources which are specific within the public health system (UMLS, SNOMED, etc.) in order to improve applications. On the other hand, the information extracted from the processing task must be used to enrich the information retrieval tools.
Keywords: information extraction, natural language technologies, lexical resources

1 Descripción General

Durante los últimos años, el campo de la biomedicina ha experimentado un desarrollo vertiginoso. Las investigaciones generan grandes volúmenes de datos biomédicos experimentales y computacionales que van acompañados de un crecimiento exponencial de las publicaciones que los describen. Esta gran cantidad de publicaciones ha superado a la mayor parte de los profesionales de la sanidad debido a que no es posible mantenerse al día de todo lo publicado sobre, por ejemplo, eventos adversos sobre fármacos. Este crecimiento continuo unido a la diversificación de la literatura biomédica requiere esfuerzos sistemáticos y automatizados que utilicen la información subyacente.

[1] MULTIMEDICA (TIN2010-20644-C03)

Como hipótesis de partida en MULTIMEDICA se tiene que es posible mejorar la búsqueda de información cuando se aplican técnicas de extracción de información (EI) que facilitan la obtención de conocimiento estructurado incluido en los textos. También es posible mejorar los módulos de recuperación de información (RI), la mejora en la construcción de índices y en los tiempos de acceso a la información, en la caracterización de los documentos empleando recursos específicos del dominio de la salud así como el uso de técnicas de aprendizaje automático y otras basadas en conocimiento explícito lingüístico (como patrones y reglas). En este proyecto se trata información en web (existente en bases de datos científicas como Medline[2] y otras más generales procedentes de ediciones digitales de periódicos y otros recursos) y terminologías propias del dominio (como UMLS[3] que incorpora varios idiomas). El proyecto tiene como finalidad definir técnicas específicas de extracción de conocimiento en el dominio de textos biomédicos con el fin de facilitar la búsqueda y la representación de esta información. Se trabaja en dos tipos de aplicaciones: la extracción de conocimiento sobre fármacos a partir de publicaciones científicas (orientada a usuarios especializados como farmacéuticos o personal clínico) y la búsqueda de información sobre salud por parte de usuarios no expertos en textos divulgativos. Así mismo, se plantea para el dominio de textos divulgativos sobre salud prestar especial atención al español, al árabe y para el dominio de textos científicos el inglés (dado que es el lenguaje mayoritario en publicación científica).

El proyecto ha comenzado en enero de 2011 y finalizará en diciembre de 2013. Participan el Grupo de Bases de Datos Avanzadas[4] de la Universidad Carlos III de Madrid, el Grupo de Sistemas Inteligentes de la Universidad Politécnica de Madrid[5] y el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid[6].

2 Objetivos

El proyecto MULTIMEDICA se plantea como retos científicos:

- Investigación en técnicas de Extracción de Información (entidades y relaciones semánticas) favoreciendo la integración de recursos terminológicos específicos del dominio de salud, en particular en español.
- Uso y mejora de corpus disponibles (tanto propios como externos).
- Tratamiento de fenómenos lingüísticos de especial interés en este dominio como son la negación, la modalidad y la anáfora.
- Mejora de las técnicas aplicadas a la extracción de relaciones semánticas, en particular de las interacciones entre fármacos, y extraer información adicional que rodea este fenómeno de especial interés como son las dosis de fármacos, efectos adversos, severidad, etc. y su integración en una ontología que permita buscar y hacer inferencia.

Desde el punto de vista de los retos tecnológicos, se trabajará en la creación de un prototipo de búsqueda en información científica otro orientado a pacientes sobre información divulgativa de salud y finalmente otro dedicado a la enseñanza de terminología específica del dominio y ayuda a la traducción.

3 Situación actual

En la actualidad se dispone de un prototipo en web para la detección de fármacos e interacciones sobre la colección de abstracts de Medline del 2010[7], basado en (Segura-Bedmar, Martínez y De Pablo-Sánchez, 2011).

También se ha lanzado el workshop DDIExtraction[8] en el marco del congreso SEPLN 2011, cuyo objetivo es reunir a los investigadores que trabajan en extracción de relaciones sobre textos biomédicos y que puedan probar sus sistemas de extracción de información en una tarea específica.

Bibliografía

Segura-Bedmar, I., Martínez, P., de Pablo-Sánchez, C.: Using a shallow linguistic kernel for drug-drug interaction extraction. Journal of Biomedical Informatics. 2011 To appear.

[2] http://www.ncbi.nlm.nih.gov/pubmed/
[3] http://www.nlm.nih.gov/research/umls/
[4] http://labda.inf.uc3m.es
[5] http://www.gsi.dit.upm.es/
[6] http://www.lllf.uam.es/
[7] http://163.117.129.57:8080/ddiextractorweb
[8] http://labda.inf.uc3m.es/DDIExtraction2011/