Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 209-212 recibido 05-05-12 revisado 29-05-12 aceptado 05-06-12

Prototipo buscador de información médica en corpus multilingües y extractor de información sobre fármacos.

Medical information search and information extraction about drugs prototype on a multilingual corpus.

Daniel Sánchez-Cisneros[1], Sara Lana[2], Antonio Moreno[3], Leonardo Campillos[3]
Paloma Martínez[1], Isabel Segura-Bedmar[1]
[1]Departamento de Informática,
Universidad Carlos III de Madrid
{dscisner, pmf, isegura}@inf.uc3m.es
[2]Universidad Politécnica de Madrid
Carretera de Valencia, Km 7
slana@diatel.upm.es
[3]Departamento de Lingüística General
Universidad Autónoma de Madrid
{antonio.msandoval, leonardo.campillos}@uam.es

Resumen: La investigación y desarrollo de nuevos fármacos ha provocado un crecimiento exponencial de la documentación relacionada con el dominio farmacológico y en la industria farmacéutica. Esto ha supuesto un problema para los profesionales del sector, debido a que tienen que invertir una gran cantidad de tiempo y esfuerzo en la revisión de esta documentación para mantener actualizados sus conocimientos. Este trabajo presenta un prototipo que busca información sobre términos médicos en colecciones divulgativas de medicina multilingües (en inglés, español, árabe y japonés) indexadas según conceptos de UMLS. El prototipo también detecta los fármacos y sus interacciones presentes en los textos.
Palabras clave: Recuperación de Información, Extracción de información.

Abstract: Research and development of new drugs has caused an exponential growth of the documentation related to drug control in the pharmaceutical industry. This is a problem for professionals since they have to invest a lot of time and effort in reviewing this documentation. This paper presents a prototype that is able to search information about medical terms over a multilingual collections of documents (English, Spanish, Arabic and Japanese), indexed with UMLS concepts. The prototype also detects drugs and drug-drug interactions discovered in the texts.
Keywords: Information Retrieval, Information Extraction.

1 Introducción

En los últimos años el tamaño de la documentación científica ha sufrido un crecimiento exponencial debido a los avances de investigación en áreas como la química, la medicina o la biología. Gran parte de esta información se almacena en grandes bases de datos bibliográficas que recopilan estos trabajos científicos. En el ámbito de la farmacovigilancia, todos los días se reportan nuevos efectos adversos, lo que supone un importante problema para los profesionales del sector de la salud, que necesitan invertir gran cantidad de tiempo y esfuerzo en la revisión de toda la documentación publicada sobre efectos adversos, y en particular, sobre interacciones farmacológicas.

En este trabajo se describe un prototipo que permite realizar búsquedas sobre distintas colecciones de documentos, y procesar cada uno de los documentos para detectar los fármacos y las interacciones farmacológicas presentes en ellos. Distintos sistemas como

ISSN 1135-5948

PubMed[1] o PIE[2] permiten realizar búsquedas sobre la base documental MedLine[3] . Nuestra principal aportación es que nuestro prototipo permite realizar búsquedas sobre distintas colecciones multilingües como MedLine, Harrison[4], tuOtroMedico[5], OcuSalud[6], etc. Además, los documentos que utilizamos han sido etiquetados con los conceptos del metatesauro UMLS (Bodenreider, 2004) que aparecen en el texto, basándonos en los recursos de diccionarios biomédicos MeSH[7] y SNOMED[8] para conceptos en inglés y en español respectivamente.

En el ámbito de la extracción de entidades y relaciones en el dominio biomédico, la mayor parte de la investigación se ha centrado en el dominio biológico. Los principales avances se han desarrollado bajo el marco del foro de evaluación BioCreative[9]. En este dominio cabe destacar la herramienta Reflect (Pafilis et al., 2006) que permite reconocer entidades como genes y proteínas, y mostrar información sobre sus interacciones en textos de páginas web. Otro ejemplo es el sistema iHOP (information Hyperlinked Over Proteins) (Hoffmann y Valencia, 2005) que tiene una colección de textos procesados para recuperar de manera sencilla información de interacciones entre proteínas.

Para fomentar el desarrollo de sistemas de extracción de información en el dominio farmacológico, el año pasado organizamos la tarea DDIExtraction (Segura-Bedmar, Martínez y Sánchez-Cisneros, 2011). La mayoría de los trabajos fueron basados en técnicas de aprendizaje automático supervisado, y en particular, los métodos kernels obtuvieron los mejores resultados F1 65%) (Thomas et al., 2011) (Mahbub et al., 2011) (Mahbub y Lavelli, 2011). Desgraciadamente ninguno de los equipos participantes desarrollo un prototipo para que los usuarios del dominio pudieran interpretar la información detectada por sus sistemas.

Por ello, nuestra principal aportación es el desarrollo de una herramienta online[10] que permita a médicos y farmacéuticos acceder de una forma más eficaz a la información relativa a fármacos y sus interacciones. Para cada uno de los fármacos detectados, la herramienta permite mostrar información procedente de distintas fuentes, como DrugBank[11] o Pubchem[12], de una forma integrada. La herramienta también muestra la lista de interacciones farmacológicas

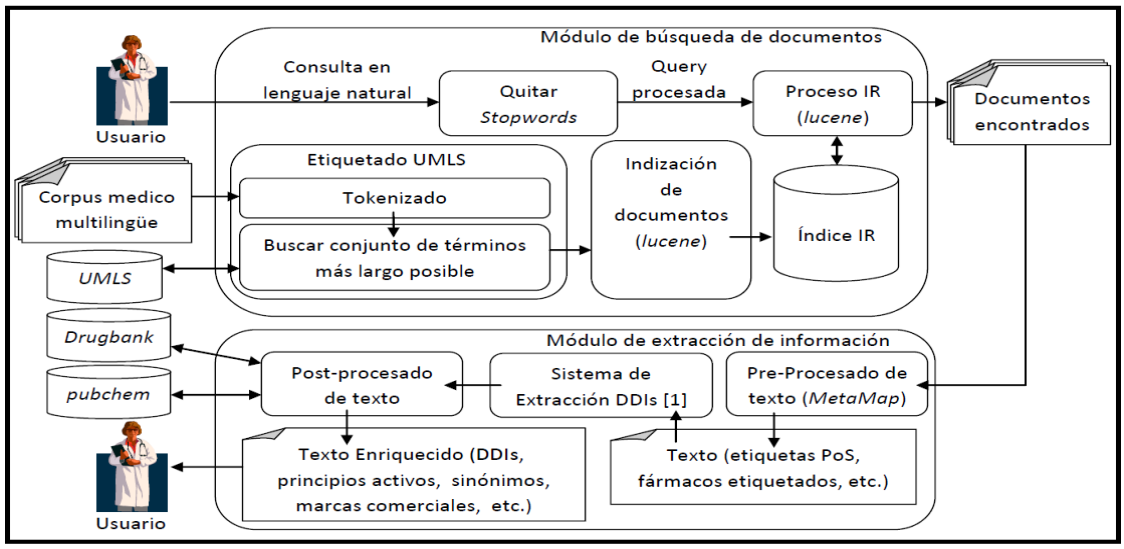

Figura 1: Arquitectura del sistema.

presentes en el texto. Estas utilidades proporcionan un mejor y más ágil acceso a la información, necesaria en la toma de decisiones respecto a la administración de un determinado fármaco.

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/PIE/
[3] http://www.nlm.nih.gov/pubs/
[4] www.harrisonmedicina.com
[5] www.tuotromedico.com
[6] http://www.ocu.org/ocu-salud-s501.htm
[7] http://www.nlm.nih.gov/pubs/
[8] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479961/
[9] http://www.biocreative.org/
[10] 163.117.129.57:8080/newddiextractorweb/
[11] http://www.drugbank.ca/
[12] http://pubchem.ncbi.nlm.nih.gov/

2 Arquitectura del sistema

El sistema está compuesto por la arquitectura general representada en la figura 1, donde se puede diferenciar un módulo de búsqueda y un módulo para la extracción de información. El módulo de búsqueda permitirá al usuario realizar búsquedas de textos biomédicos en un repositorio de colecciones de documentos multilingües. Por otro lado, el segundo módulo permite procesar textos para detectar fármacos y extraer sus interacciones farmacológicas, así como obtener información extendida de cada fármaco identificado en el texto.

2.1 Módulo de Búsqueda.

En primer lugar, el sistema permite buscar documentos en un repositorio multilingüe. Este repositorio ha sido creado por el grupo LLI de la universidad Autónoma de Madrid[13] y contiene colecciones de documentos en los siguientes idiomas:

- Inglés: más de 19.000.000 documentos en inglés de la colección biomédica Medline 2010.
- Español: 4.204 documentos en español de las revistas biomédicas Harrison, OcuSalud y TuOtroMedico.
- Japonés: 4.746 documentos en japonés de un revistas médicas de especialidades variadas.
- Árabe: 43.526 documentos en árabe del portal médico Altibbi[14].

Esto hace un total de 19.051.476 documentos de carácter biomédico, sin embargo la mayoría (19 millones) están en inglés. Para realizar el módulo de búsqueda se ha usado la herramienta IR Apache Lucene[15].

Las colecciones en inglés y español han sido procesadas por el grupo GSI de la Universidad Politécnica de Madrid[16] para etiquetar todos los términos MeSH y SNOMED identificados en los textos. MeSH es un tesauro de términos en inglés que es utilizado para indexar los artículos de MedLine. SNOMED es una terminología clínica que permite representar la información clínica de forma multilingüe.

En este proceso de etiquetado se han tenido que realizar tareas de desambiguación, ya que un término puede tener varias acepciones. Para tratar este problema se ha seguido la siguiente estrategia de prioridades:

1. Buscar los conceptos de MeSH y SNOMED que contengan todas las palabras del término.
2. Buscar los conceptos de MeSH y SNOMED que contengan el conjunto mayor de palabras del término.
3. Buscar los conceptos de MeSH y SNOMED que contengan alguna de las palabras del término.

Con este conjunto de corpus etiquetado con etiquetas MeSH y SNOMED hemos creado un índice sobre el que realizaremos los procesos de búsqueda en nuestra herramienta. Esto permite al usuario realizar búsquedas más avanzadas por conceptos.

Como resultado de estas búsquedas se devuelve un listado de documentos facilitando su identificador de Medline (PMID), titulo, revista, autores, snippet, etc. Finalmente el usuario puede ver el texto de cada documento, así como procesar el texto en busca de relaciones semánticas (interacciones entre fármacos).

2.2 Módulo de Extracción de Información.

El módulo de extracción permite procesar textos, ya sean resultados del módulo de búsqueda o textos introducidos directamente por el usuario. Para ello, en este módulo se ha realizado un trabajo de integración de varios recursos:

- En primer lugar, los textos son analizados semánticamente por MetaMap[17] para identificar los fármacos.
- A continuación, los textos son procesados por el sistema DrugDDI (Segura-Bedmar, 2010) que permite la extracción de interacciones farmacológicas. Dicho sistema está basado en el Shallow Linguistic Kernel (Giuliano, Lavelli y Romano, 2006).
- Finalmente, el sistema busca información para cada fármaco detectado: el nombre de su principio activo, código ATC, nombres comerciales, descripción del fármaco, etc. Para obtener toda esta información, utilizamos bases de datos

[13] http://www.lllf.uam.es/ESP/
[14] http://www.altibbi.com/
[15] http://lucene.apache.org/
[16] http://www.gsi.dit.upm.es/
[17] http://metamap.nlm.nih.gov/

farmacológicas como Drugbank y Pubchem.

Como resultado de este proceso, el sistema devuelve por un lado el texto con los fármacos identificados e información adicional de cada fármaco, y por otro lado un listado de las interacciones identificadas en el texto (ver figura 2).

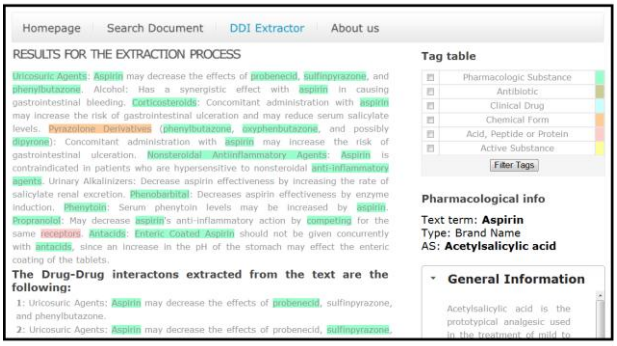

Figura 2: Ejemplo de procesado de texto.

3 *Trabajo en curso.*

El sistema permite la búsqueda de documentos sobre determinados conceptos clínicos (enfermedades, fármacos, síntomas, etc.), y además, el sistema es capaz de procesar los documentos para extraer de forma dinámica y online las interacciones entre fármacos que se describen en el texto. Para ello, procesa información estructurada y no estructurada.

Para ello, se ha incorporado un repositorio multilingüe de colecciones de documentos en diferentes idiomas, al que se ha realizado un etiquetado para identificar los términos MeSH y SNOMED en inglés y español, utilizando sus diccionarios de conceptos médicos.

Como trabajo futuro nos planteamos investigar en la detección de la gravedad de la interacción, su grado de certeza, y otros factores que pueden influir en la interacción como la dosis, tiempo de ingesta entre medicamentos, características individuales del paciente, etc. Toda esta información es vital a la hora de determinar la importancia clínica de una determinada interacción y que el facultativo sea capaz de tomar la decisión correcta respecto a la administración de un determinado fármaco.

4 *Agradecimientos.*

Este trabajo ha sido desarrollado en el marco del proyecto MA2VICMR (S2009/TIC-1542) y MULTIMEDICA[18] (TIN2010-20644-C03-01).

[18] http://labda.inf.uc3m.es/multimedica/

Bibliografía

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res., 267-270.

Giuliano, C., Lavelli, A., Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. EACL-2006, (págs. 401 - 408).

Hoffmann, R., Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics.

Mahbub, M., Ben, A., Lavelli, A., y Zweigenbaum, P. (2011). Two Different Machine Learning Techniques for Drug-Drug Interaction Extraction . DDIExtraction2011: First Challenge Task on Drug-Drug Interaction Extraction 2011, 19 - 26.

Mahbub, M.F., y Lavelli, A. (2011). Drug-drug Interaction Extraction Using Composite Kernels. DDIExtraction2011: First Challenge Task on Drug-Drug Interaction Extraction 2011, 27 - 33.

Pafilis, E., O'Donoghue, S.I., Jensen, L.J., Horn, H., Kuhn, M., Brown, N.P., y Schneider, R. (2006). Reflect: augmented browsing for the life scientist. Nature Biotechnology, 508-510.

Segura-Bedmar, I. (2010). Application of information extraction techniques to pharmacological domain: extracting drug-drug interactions. Universidad Carlos III de Madrid, Departamento de Informática.

Segura-Bedmar, I., Martínez, P. Sánchez-Cisneros, D., (2011). The 1st DDIExtraction-2011 Challenge Task: Extraction of Drug-Drug Interactions from Biomedical Texts. SEPLN 2011, (págs. 1 - 9). Huelva.

Thomas, P., Neves, M., Solt, I., Tikk, D., y Leser, U. (2011). Relation Extraction for Drug-Drug Interactions using Ensemble Learning . DDIExtraction 2011: First Challenge Task on Drug-Drug Interaction Extraction 2011., 11 - 18.