Actas V Jornadas TIMM, pp 37-39

Propuesta de un sistema de extracción de información farmacoterapéutica a partir de documentos especializados procedentes de diversas fuentes en castellano

Isabel Moreno

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
imoreno@dlsi.ua.es

M.T. Romá-Ferri

Departamento de Enfermería
Universidad de Alicante
mtr.ferri@ua.es

Paloma Moreda

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
moreda@dlsi.ua.es

Resumen

Hoy en día, disponemos de una gran cantidad de información digital relativa a la salud. El uso de esta información, mayoritariamente textual, resulta crítico para innovar en las investigaciones médicas, para mejorar la calidad de la atención sanitaria y para reducir costes [FRC13]. Y sin embargo, el personal sanitario tiene dificultades para poder aprovechar tal cantidad de información multilingüe dispersa en múltiples fuentes de información.

En la actualidad, los esfuerzos se centran, sobre todo, en crear herramientas y recursos para lengua inglesa. Esto se traduce en carencias para los profesionales sanitarios en países de habla no inglesa. Por ello, el objetivo de este proyecto de tesis doctoral es analizar y proponer nuevas técnicas y enfoques que permitan abordar la creación de un sistema de extracción de información farmacoterapéutica a partir de documentos especializados procedentes de diversas fuentes en castellano.

In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

1. Introducción

Hoy en día tenemos a nuestra disposición una gran cantidad de información digital. Lo mismo le ocurre al personal sanitario durante su actividad profesional.

Un ejemplo evidente de la gran cantidad de información disponible, son las bases de datos científicas como MEDLINE[1], con las ultimas novedades sanitarias generalmente en inglés.

A la información científica hay que añadir la generada en la Historia Clínica Electrónica (HCE) de los pacientes, en la lengua nativa del profesional. En la HCE se pueden encontrar muchos campos con información textual libre, generalmente en la lengua nativa del profesional. Sobre todo encontramos información textual en aquellos campos relacionados con la medicación como la posología y las indicaciones tanto al paciente como al farmacéutico.

Emplear toda esta información resulta crítico para innovar en las investigaciones médicas, para mejorar la calidad de la atención sanitaria y para reducir costes [FRC13]. Hoy en día la mayoría de los esfuerzos se centran en crear herramientas y recursos lingüísticos (necesarios para construir o evaluar estas herramientas) en lengua inglesa. Esto se traduce en una carencia de herramientas y recursos para lengua nativa del profesional en países de habla no inglesa.

2. Propuesta

Por todo lo expuesto anteriormente, el objetivo final de este proyecto de tesis doctoral consiste en analizar y proponer nuevas técnicas y enfoques que permitan la construcción de un sistema de Extracción de Información (EI) farmacoterapéutica en castellano. Con ello se convertirá la información textual de documentos especializados en información estructurada. Lo que permitirá presentarla de forma organizada, facilitando su consulta e interpretación en el menor tiempo posible.

Para conseguir nuestro objetivo final, se plantean a su vez 4 objetivos:

- Definir un esquema de anotación semántico: Se ha definido un esquema de anotación semántico compuesto por 18 elementos farmacoterapéuticos. Dichos elementos están basados en las necesidades de los profesionales sanitarios y de los pacientes, así como en trabajos de referencia en este dominio: corpus i2b2[USXC10], en los sistemas de [DGZ10, PC10] y en la ontología farmacoterapéutica, OntoFIS[RF09]. En el cuadro 1 se encuentran todos los elementos de nuestro esquema, tanto entidades nombradas como relaciones entre las mismas.

Cuadro 1: El esquema de anotación propuesto

Medicament (Medicamento)	Disease (Proceso clínico)
Drug (Principio Activo)	Desirable Effect (Efecto deseado)
Chemical Composition (Composición Química)	Therapeutic Indication (Indicación Terapéutica)
Route (Vía de administración)	Therapeutic Action (Acción terapéutica)
Pharmaceutical Form (Forma farmacéutica)	Side Effect (Efecto secundario)
Food (Alimento)	Unit Of Measurement (Unidad de medida)
Infectious Agent (Agente Infeccioso)	Contraindication (Contraindicación)
Toxic Agente (Agente Tóxico)	Overdosage (Sobredosis)
Excipient (Excipiente)	Interaction (Interacción)

- Anotar semánticamente un corpus de documentos especializados: Se ha anotado semánticamente un corpus de documentos especializados con nuestro esquema. En concreto, hemos usado fichas técnicas de medicamento, que son una versión ampliada de los prospectos de los medicamentos para los profesionales sanitarios. La Figura 1 muestra un fragmento de ficha técnica anotada con algunos de nuestros conceptos como indicación terapéutica o principio activo.

[1] Con más de 23 millones de documentos indizados actualmente (julio 2014): http://www.ncbi.nlm.nih.gov/pubmed

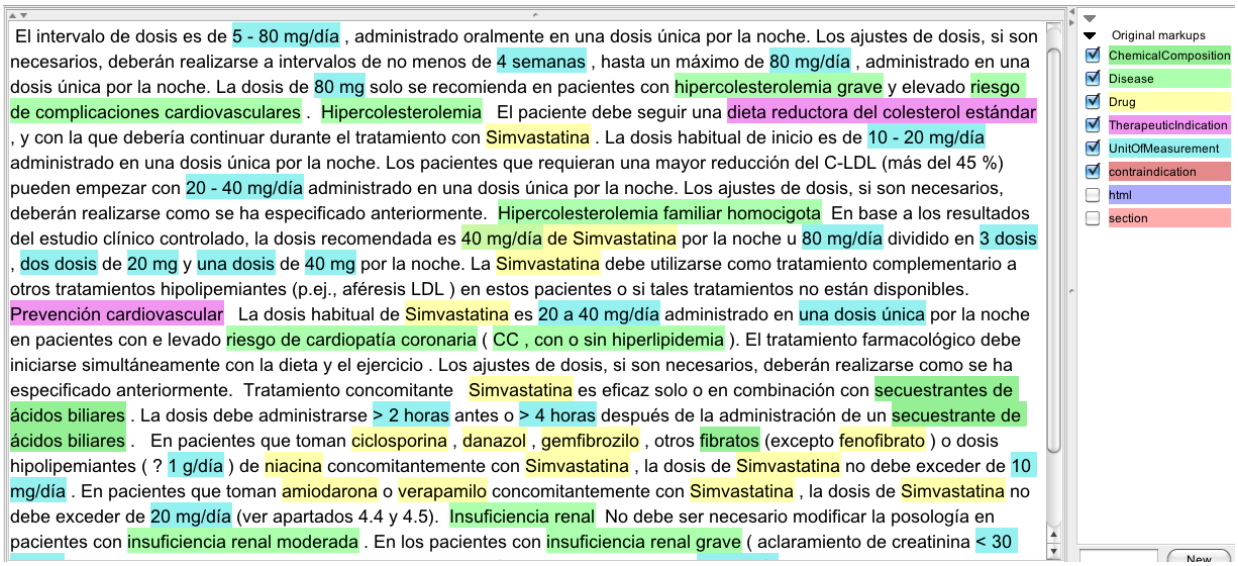

Figura 1: Ejemplo de texto marcado con varios conceptos del esquema propuesto

- Crear o adaptar recursos lingüísticos y semánticos: Se ha analizado como adaptar recursos semánticos como la ontología OntoFIS[RF09] y la terminología sanitaria Snomed[CR80] y así enriquecer nuestro sistema de Extracción de información. En un futuro estudiaremos y crearemos otros recursos que podamos necesitar.
- Crear el sistema de EI empleando distintas heurísticas. Hemos realizado algunas pruebas con sistemas basados en diccionarios y los recursos comentados en el objetivo anterior. Según las necesidades de cada tipo de elemento estudiaremos si utilizar reglas o aprendizaje automático o un enfoque hibrido para su extracción.

Referencias

[CR80] R.A. Cote and S. Robboy. Progress in medical information management: the systematized nomenclature of medicine (snomed) [progres dans la gestion de l'information medicale. la nomenclature systematisee de la medecine (snomed)]. *Union Medicale du Canada*, 109(9):1243–1252, 1980.

[DGZ10] Louise Deléger, Cyril Grouin, and Pierre Zweigenbaum. Extracting medication information from French clinical texts. *Studies in health technology and informatics*, 160(Pt 2):949–53, January 2010.

[FRC13] Carol Friedman, Thomas C Rindflesch, and Milton Corn. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of biomedical informatics*, 46(5):765–73, October 2013.

[PC10] Jyotishman Pathak and Christopher G Chute. Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. *Journal of the American Medical Informatics Association : JAMIA*, 17(4):432–9, January 2010.

[RF09] M.T. Romá-Ferri. *OntoFIS: tecnología ontológica en el dominio farmacoterapéutico*. PhD thesis, Universidad de Alicante, 2009.

[USXC10] Ozlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association : JAMIA*, 17(5):519–23, 2010.