Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 157-160 recibido 23-03-2015 revisado 27-04-2015 aceptado 14-05-2015

# EXTracción de RElaciones entre Conceptos Médicos en fuentes de información heterogéneas (EXTRECM)*

## *EXTracción de RElaciones entre Conceptos Médicos*

**Arantza Díaz de Ilarraza†**
**Koldo Gojenola‡**
UPV/EHU
†Paseo Manuel Lardizabal, 1, 20018 San Sebastián
‡Paseo Rafael Moreno Pitxitxi, 3, 48013 Bilbao
a.diazdeilarraza,koldo.gojenola@ehu.eus

**Lourdes Araujo**
**Raquel Martínez**
UNED
C/ Juan del Rosal, 16. 28040 Madrid
lurdes,raquel@lsi.uned.es

**Resumen:** En este proyecto se plantea la extracción de relaciones entre conceptos médicos en documentos científicos, historiales médicos e información de carácter general en Internet, en varias lenguas utilizando técnicas y herramientas de Procesamiento de Lenguaje Natural y Recuperación de Información. El proyecto se propone demostrar, mediante dos casos de uso, los beneficios de la aplicación de este tipo de tecnologías lingüísticas al dominio de la salud.
**Palabras clave:** identificación relaciones entre conceptos, dominio médico, minería de textos

**Abstract:** This project addresses extraction of medical concepts relationship in scientific documents, medical records and general information on the Internet, in several languages by using advanced Natural Language Processing and Information Retrieval techniques and tools. The project aims to show, through two use cases, the benefits of the application of language technology in the health sector.
**Keywords:** identification of concept relationship, medical domain, text mining

## *1 Descripción general*

El proyecto EXTRECM (http://ixa.si.ehu.es/extrecm) tiene como objetivo principal proporcionar un acceso eficiente y fiable al gran volumen de información al que en este momento acceden los profesionales de la salud de manera manual o casi artesanal. Este volumen no solo corresponde a documentos científicos alojados en repositorios específicos, sino que consideramos, además, que la información contenida en la web sobre páginas especializadas y/o redes sociales puede aportar información de distinta naturaleza basada en la experiencia de los pacientes, que puede complementar a las otras fuentes. Es importante que estos profesionales puedan disponer de mecanismos que les faciliten el "acceso avanzado" a la información contenida en todos estos millones de documentos de naturaleza heterogénea. Por "acceso avanzado" entendemos un acceso que permita concentrarse en el concepto médico deseado y recuperar la información relacionada con dicho concepto médico presente en los diferentes documentos y fuentes de información heterogéneas. En este proyecto nos planteamos precisamente el reto de desarrollar y aplicar las tecnologías de tratamiento del lenguaje a diversos tipos de documentos que manejan los profesionales del área de la salud en múltiples idiomas y a escala web.

Los profesionales en el sector de la salud pública tienen que acceder a conocimiento preciso y completo para poder tomar decisiones con la mayor cantidad de información posible. Cada vez es más difícil tomar estas decisiones dado el gran volumen de datos que ha de considerarse. Este volumen dificulta encontrar manualmente relaciones que pueden ser utilizadas en la extración de conoci-

* TIN2013-46616-C2-1-R, TIN2013-46616-C2-2-R

ISSN 1135-5948 

miento. Este proyecto se centra en tres tipos de colecciones de documentos: publicaciones científicas, historiales clínicos e información de carácter general de la web, redes sociales, blogs de usuarios, etc. Las redes sociales, y particularmente Twitter, permiten introducir las valoraciones de los pacientes y allegados con respecto a una enfermedad, tratamiento, medicamento, etc. que normalmente quedan fuera de las fuentes de consulta habituales de los profesionales de la salud.

Los profesionales médicos del área de la salud están habituados a realizar consultas a algunos de estos tipos de documentación, aunque normalmente se limitan a búsquedas por palabras clave. En este proyecto se le da una nueva perspectiva a estos profesionales que están inmersos en una constante carrera para estar informados y responder adecuadamente a cualquier cambio, desarrollo o novedad. Por este motivo es importante disponer de tecnología que filtre, seleccione y organice dicha información.

Las técnicas de PLN (Procesamiento del Lenguaje Natural) y RI (Recuperación de Información) nos permitirán crear un sistema de vigilancia tecnológica, tanto de novedades científicas de interés para los expertos, como de preocupaciones e intereses sociales relacionados con la salud y reflejados en las redes. Esta vigilancia iría más allá de la información aportada por una búsqueda clásica, ya que incluiría relaciones indirectas entre los conceptos involucrados. En este momento no existe ningún sistema de consulta avanzada sobre términos/conceptos médicos en el cual el experto en medicina (doctores u otro personal sanitario) pueda formular su pregunta de forma "dirigida" por el sistema y además en inglés, español o euskera. Tampoco ningún recurso que sea capaz de realizar búsquedas en fuentes de información heterogéneas: historiales clínicos, publicaciones científicas, redes sociales e Internet en general. El proyecto EXTRECM supone una innovación que permitirá obtener las respuestas partiendo de repositorios médicos muy extensos usados por la comunidad médica internacional. Así, ayudará a eliminar posibles barreras idiomáticas y pondrá al alcance del personal sanitario toda la información existente en los mencionados repositorios.

## *2 Grupos involucrados*

El proyecto tiene una naturaleza multidisciplinar y será abordado mediante la colaboración entre grupos de investigación expertos en tecnologías de la lengua y del área de la salud. Esta colaboración puede ayudar a la creación de sinergias entre las dos partes, con el objetivo principal de crear herramientas de procesamiento de textos médicos que mejoren la eficiencia y competitividad de los sistemas de salud y hospitalarios, posibilitando el acceso a ingentes cantidades de información.

Los grupos implicados en el proyecto son:

- Grupo IXA[1] la Universidad del País Vasco UPV/EHU. Tiene una amplia trayectoria en investigación en Procesamiento de Lenguaje Natural y lingüística computacional, y de participación en proyectos de investigación. Tiene líneas de investigación abiertas en el dominio médico.
- Grupo NLP&IR[2] de la UNED. Dispone de una amplia experiencia en Acceso Inteligente a la Información y Adquisición y Representación de Conocimiento Léxico, Gramatical y semántico. Tiene una amplia trayectoria en la realización de proyectos de investigación.
- Hospitales de Galdakao (HGA) y Basurto (HUB), integrados en el grupo de trabajo IXA pertenecientes al Servicio Público de Salud. Este grupo es pionero en el tratamiento e implantación de los historiales clínicos electrónicos, siendo un socio fundamental en este proyecto. Aportará su experiencia en el área de la detección de efectos adversos manifestados explícita o implícitamente en los historiales clínicos (caso de uso).
- Orphanet[3], entidad internacional dedicada al objetivo de contribuir a la mejora del diagnóstico, cuidado y tratamiento de los pacientes con enfermedades raras. En este proyecto esta entidad se integra en el grupo de trabajo de la UNED. Ellos nos aportan su conocimiento y necesidades en el escenario de recuperación de información para enfermedades raras (caso de uso).

[1] `http://ixa.si.ehu.es/Ixade`
[2] `http://nlp.uned.es/`
[3] `http://www.orpha.net/`

### 2.1 Casos de uso

Las técnicas propuestas se aplicarán a dos casos de uso específicos de interés para las instituciones médicas que colaboran en el proyecto: los hospitales de Galdakao y Basurto para el caso de identificar efectos adversos (EA) a medicamentos, y Orphanet para el caso de asociar discapacidades a enfermedades raras (EERR).

Los grupos IXA y UNED colaboran en la construcción de herramientas de PLN para el dominio de la salud en un entorno multilingüe. Esas herramientas se utilizan y se ponen a prueba en tipos de documentos heterogéneos.

IXA aplica esas herramientas de PLN en su colaboración con los hospitales de Galdakao y Basurto, para quienes identifica posibles reacciones adversas a medicamentos en informes médicos y después en los documentos recuperados de Internet. Por su parte, UNED aplica esas herramientas a distintos tipos de documentos recuperados de Internet, identificando posibles discapacidades asociadas a enfermedades raras, que interesan a Orphanet.

Además, los distintos tipos de documentos tratados en ambos grupos y las técnicas adaptadas a ellos se generalizan abordando los casos de uso de interés del proyecto de forma cruzada. A partir de una selección de medicamentos indicados por el grupo IXA, el grupo UNED aplicará las técnicas desarrolladas para la búsqueda de reacciones adversas a esa selección de medicamentos. Por su parte el grupo IXA aplicará las técnicas desarrolladas para identificar casos de EERR en los informes de que dispone, de manera que puedan relacionar reacciones adversas a medicamentos y discapacidades.

El trabajo conjunto de los dos grupos de investigadores, junto con los investigadores incluidos en los correspondientes equipos de trabajo, en su mayoría profesionales de la salud, se considera un motor generador de nuevas ideas, que podrán ser puestas en práctica en el mundo de la salud.

## *3 Objetivos*

En la figura 1 se muestra la interrelación entre los principales objetivos del proyecto. Tomando como base el estado del arte en el área, se trata de definir los requerimientos del usuario sobre el tipo de consultas avanzadas a grandes volúmenes de información que se desean realizar en el dominio médico en los tres grandes bloques de documentación con los que nos planteamos trabajar: historiales clínicos, web sociales y artículos médicos.

Desde el punto de vista cuantitativo, es importante identificar los volúmenes de datos con los que tienen que trabajar nuestros expertos. Desde el punto de vista cualitativo, es importante estudiar la estructura de los documentos con los que se va a trabajar, formatos, tipos de información que se maneja, etc. Hemos de observar el modo de trabajo de los expertos incluidos en los grupos de trabajo de cada subproyecto ya que como resultado de esa observación conoceremos los requisitos que tienen que cumplir los sistemas que se desarrollen con el objetivo de servir de la manera más precisa a sus necesidades.

También es necesario seleccionar y preparar y, en su caso, etiquetar el conjunto de documentos de referencia para la evaluación de los resultados para los idiomas inglés, castellano y euskera. Las exigencias son diferentes para cada uno de los idiomas y el trabajo de etiquetado será más exigente para los documentos en castellano que para los escritos en euskera. Este paso va ligado al diseño e implementación de los módulos de acceso, recuperación, filtrado y organización de la información relacionados con los casos de uso.

Otra parte fundamental del proyecto es la preparación, diseño e implementación de los módulos de procesamiento del lenguaje. La idea general del proyecto es utilizar una arquitectura abierta. Son necesarios procesadores básicos para todas las lenguas del proyecto tales como tokenización, lematización y etiquetado morfosintáctico, análisis sintáctico, desambigüación semántica, y reconocimiento de entidades nombradas. Algunos de estos procesadores pueden ser módulos genéricos de procesamiento del lenguaje que han de adaptarse al dominio de la salud, pero hay otro grupo importante de herramientas que tienen que desarrollarse expresamente para cumplir los objetivos de este proyecto.

Utilizando las técnicas y herramientas mencionadas se abordará la construcción de prototipos para tratar los dos casos de uso considerados: detección de eventos adversos a medicamentos, y discapacidades asociadas a efectos adversos, en los distintos tipos de documentos considerados. Nos proponemos investigar nuevas técnicas para la detección de las conexiones más relevantes entre los con-

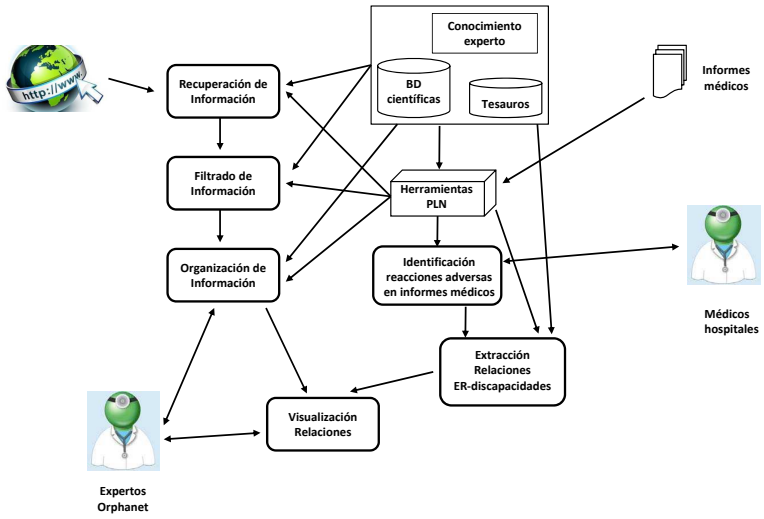


Figura 1: Relación entre los componentes del proyecto: los hospitales colaboradores y distintas fuentes de Internet proporcionan información que se procesa con técnicas de PLN desarrolladas entre ambos grupos, IXA y UNED. Orphanet y los hospitales contribuyen a la evaluación de los resultados obtenidos.

ceptos considerados. Estas conexiones pueden interpretarse como una asociación entre los conceptos implicados, que permitirá la generación de conocimiento y la confirmación de relaciones procedentes de otras fuentes. Para cumplir este objetivo se necesita por una parte identificar este tipo de conceptos mediante el uso de ontologías de dominio, reconocimiento de entidades nombradas, patrones, sinónimos, etc. y por otra parte aplicar métodos que permitan seleccionar las relaciones realmente significativas y presentarlas de forma accesible a los profesionales de la salud.

Otra fase fundamental del proyecto es la evaluación de los casos de uso con expertos. Y finalmente abordaremos la generalización de los resultados obtenidos en el proyecto. Se trata de analizar la generalidad de las técnicas desarrolladas y fomentar la interacción entre los grupos. Para comprobar esta hipótesis aplicaremos los sistemas desarrollados a los casos de uso cruzados. Así, el grupo IXA proporcionará al grupo UNED una serie de casos de reacciones adversas a medicamentos encontrados en los informes médicos. Por su parte, el grupo UNED aplicará las técnicas desarrolladas para buscar información en documentación científica y redes sociales que confirme esta hipótesis y pueda aportar detalles adicionales a esta información. Para ello las relaciones que se considerarán en este caso son medicamentos y reacciones adversas.

## *4 Situación Actual*

El proyecto está en una fase inicial ya que lleva pocos meses activo. En relación a las colecciones que se van a utilizar en el proyecto, por una parte se está trabajando en la ampliación de los corpus anotados manualmente con efectos adversos a medicamentos, y en la mejora del asistente de anotación manual. Por otra parte se están compilando diferentes corpus con información sobre enfermedades raras y las posibles discapacidades asociadas. Esta información abarca la web y los artículos científicos.

De cara a la identificación de relaciones entre medicamentos y efectos adversos se están aplicando tanto técnicas supervisadas como no supervisadas. Además se está ampliando la cobertura del anotador morfosintáctico de conceptos médicos con nuevas abreviaturas y acrónimos. También se está trabajando en la anotación automática de discapacidades y en el filtrado de documentos relevantes.