Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 133-136 recibido 29-03-2017 revisado 12-05-2017 aceptado 29-05-2017

PROcesamiento Semántico textual Avanzado para la detección de diagnósticos, procedimientos, otros conceptos y sus relaciones en informes MEDicos (PROSA-MED)

Advanced semantic textual processing for the detection of diagnostic codes, procedures, concepts and their relationships in health records

Arantza Díaz de Ilarraza[(1)], Koldo Gojenola[(2)], Raquel Martínez[(3)], Víctor Fresno[(3)], Jordi Turmo[(4)], Lluis Padró[(4)]
(1) P. M. Lardizabal, 1, 20018 San Sebastián UPV/EHU
(2) P. Rafael Moreno, 3, 48013 Bilbao UPV/EHU
(3) C/ Juan del Rosal, 16 28040 Madrid UNED
(4) C/ Jordi Girona, 1-3 08034 Barcelona UPC
koldo.gojenola@ehu.eus

Resumen: El objetivo de este proyecto es desarrollar procesadores para el análisis automático de textos médicos, poniendo a disposición de la comunidad científica y empresarial un conjunto amplio y versátil de herramientas y recursos lingüísticos para el análisis morfológico, sintáctico y semántico, así como la asignación de códigos diagnósticos y procedimientos a informes médicos según el estándar CIE-10 y la detección de relaciones entre conceptos. Se desarrollarán herramientas para el español, dado su amplio uso en sistemas de salud a nivel internacional, explorando además otras lenguas con diferentes características como el catalán y el vasco.
Palabras clave: procesamiento textos clínicos, aprendizaje automático, extracción relaciones, grafos semánticos.

Abstract: The main aim of this project will be to develop a set of processors for the automatic analysis of medical texts. The project will create a wide and exible set of tools, linguistic, and semantic resources for the following tasks: morphologic, syntactic and semantic analysis adapted to medical texts; assignment of diagnostics and procodures following the ICD-10 coding, and detection of relationships between concepts. The project will develop tools for Spanish, used in multiple health systems of different countries. Moreover, we will also tackle other languages with different characteristics such as Catalan and Basque.
Keywords: clinical text processing, machine learning, relation extraction, semantic graphs.

1 *Descripción general*

El proyecto PROSA-MED[1] es un proyecto financiado por el Ministerio de Economía, Industria y Competitividad en la convocatoria 2016 de Proyectos I+D+I, dentro del Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016. PROSA-MED se propone como continuación del trabajo realizado en el ya finalizado proyecto EXTRECM (Díaz et al., 2015).

El sector sanitario constituye un sector de vital importancia, tanto por su papel en el estado del bienestar como por su carácter multidisciplinar. El número de documentos del dominio médico generados por los centros de atención al paciente (hospitales y atención primaria) aumenta constantemente, y en ellos el desarrollo de herramientas automáticas de análisis textual puede suponer un avance crucial para los sistemas de salud. Las tecnologías de la lengua disponen de herramientas para realizar un análisis textual que ayude al personal médico a aumentar su productividad, redundando en el beneficio de todos.

[1] http://ixa2.si.ehu.eus/prosamed/

ISSN 1135-5948

El consorcio de grupos de investigación de las universidades e instituciones del ámbito sanitario que formamos parte de este proyecto estamos convencidos de la factibilidad de realizar un importante salto tecnológico en este campo. Nuestro objetivo es proponer soluciones en el tratamiento de Informes Clínicos Hospitalarios (ICH) e Historia Clínica Electrónica (HCE) a procesos que, en la actualidad, suponen un gran coste personal y económico.

En este proyecto se desarrollará un conjunto de procesadores que permitirán el análisis automático de textos médicos teniendo en cuenta criterios de robustez, alta precisión y cobertura. El proyecto pondrá a disposición del personal médico un conjunto amplio y versátil de herramientas, recursos lingüísticos, terminológicos y semánticos, que se aplicarán al tratamiento de los tipos de texto mencionados para las siguientes tareas:

- Análisis morfológico, sintáctico y semántico adaptado a textos médicos de acuerdo al estado del arte en el área, y haciendo especial énfasis en el reconocimiento de entidades.
- Asignación de códigos diagnósticos y de procedimientos a informes médicos según la especificación CIE-10 (World Health Organization, 2009).
- Detección de relaciones entre conceptos como paso previo para avanzar en el área del descubrimiento de evidencias no explícitamente expresadas en los textos.

En el proyecto se desarrollarán herramientas para distintas lenguas. El español constituye un objetivo ambicioso, dado su amplio uso en los sistemas de salud de multitud de países. Además, se explorarán otras lenguas con diferentes características y grados de desarrollo en el ámbito médico: el catalán y el vasco. El trabajo desarrollado en este proyecto tiene un gran interés en el entorno empresarial público y privado, ya que se proporcionarán soluciones software que estarán disponibles para PYMES u otras empresas que tengan interés en desarrollar productos en el dominio médico. Las entidades participantes representan a tres sistemas de salud públicos (Cataluña, Madrid y País Vasco) pero podrá extenderse a otros ámbitos y áreas de aplicación.

Esperamos que el impacto científico de este proyecto se aproveche en el contexto de la mejora general de la asistencia sanitaria, la facturación a mutuas privadas por los servicios públicos, así como en la optimización y organización global de recursos sanitarios. Asimismo, los resultados del proyecto ayudarán a resolver retos actuales como son el reconocimiento de patrones que rigen la relación entre el consumo de recursos y la actividad realizada, o determinar si existe una anomalía en la calidad de la prestación de una asistencia o el coste asociado a la misma. Asimismo, facilitará el tratamiento inteligente de las HCE y ayudará a implementar políticas de salud más eficientes, inteligentes, personalizadas y adaptadas a los pacientes, contribuyendo así a la mejora y sostenibilidad del sistema. Se espera que los resultados del proyecto puedan aplicarse directamente en el ámbito estatal, así como ser exportados a otros países hispanohablantes y adaptarse a otras lenguas. Además, dada la experiencia de los grupos de investigación participantes, se espera que este proyecto genere también un importante impacto científico en forma de publicaciones, generando nuevo conocimiento que supondrá un avance en las diferentes áreas científicas involucradas.

2 Grupos involucrados

El proyecto tiene una naturaleza multidisciplinar y será abordado mediante la colaboración entre los tres grupos de investigación participantes, expertos en tecnologías de la lengua y su aplicación al área de la salud.

PROSA-MED consta de tres subproyectos:

- IXA-MED: Técnicas supervisadas para asignación de diagnósticos CIE-10 y detección de efectos adversos.
- MAMTRA-MED: Modelado y AutoMatización de exTracción de Relaciones y cAtegorización de informes MEDicos para la recomendación de códigos CIE-10.
- GRAPH-MED: Extracción de grafos semánticos a partir de historiales clínicos textuales.

Los grupos implicados en este proyecto coordinado son:

- Grupo IXA[2] de la Universidad del País

[2] http://ixa.si.ehu.es/Ixade

Vasco UPV/EHU. Tiene una amplia trayectoria en investigación en Procesamiento de Lenguaje Natural (PLN) y Lingüística Computacional, y de participación en proyectos de investigación, con líneas de investigación abiertas en el dominio médico.

- Grupo NLP&IR[3] 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 y líneas de investigación abiertas en el dominio médico.
- Grupo TALP[4] de la UPC, con amplio historial de proyectos de investigación en Procesamiento de Lenguaje Natural y Minería de Texto. Actualmente tiene líneas abiertas de investigación en el dominio médico.
- Hospitales de Galdakao (HGA) y Basurto (HUB), integrados en el grupo de trabajo IXA pertenecientes al Servicio Público de Salud.
- Hospital Fundación Universitaria Fundación Alcorcón (HUFA). Es un hospital general, integrado en la red sanitaria publica del Servicio Madrileño de Salud y ubicado en la zona sur de la Comunidad de Madrid. Participa en el proyecto integrado en el grupo UNED.
- Fundación IDIAP Jordi Gol, integrada en el grupo TALP. IDIAP desarrolla y gestiona la investigación de la Atención Primaria de Salud principalmente en Cataluña, facilitando la participación de investigadores de sectores.

3 Objetivos

El objetivo general del proyecto PROSA-MED es proponer soluciones en el tratamiento de Informes Clínicos Hospitalarios e Historia Clínica Electrónica a procesos que, en la actualidad, suponen un gran coste personal y económico. Este objetivo general se puede concretar en los siguientes objetivos parciales:

- Desarrollar y adaptar herramientas de PLN al dominio médico. El procesamiento masivo de documentos médicos abre un abanico de opciones que puede facilitar múltiples iniciativas innovadoras con posibilidades aún desconocidas. La disponibilidad de herramientas robustas y precisas para este dominio supondrá un gran salto cualitativo, al poner a disposición de entidades, tanto públicas como privadas, estas herramientas básicas de procesamiento del dominio médico.
- Estudiar diferentes enfoques supervisados y no supervisados para la codificación automática de códigos CIE-10 en informes médicos. Una gran parte de los sistemas de salud ha empezado a codificar los diagnósticos médicos haciendo uso del CIE-10 a partir de enero de 2016, lo que supone que éste es un momento idóneo para el desarrollo de herramientas automáticas que realicen esta codificación. El proceso de asignación de un diagnóstico desde un texto se encuentra lejos de ser trivial, ya que los informes médicos están escritos en lenguaje natural y sujetos a la variabilidad inherente al lenguaje libre, como el uso de lenguaje no estandarizado. Además, el catálogo CIE-10 contiene miles de diagnósticos y procedimientos, y su detección supone un problema enormemente complejo. Este objetivo, además de suponer un gran reto científico, tiene una aplicación inmediata al proceso de informes médicos.
- Aplicación de técnicas de PLN al problema de identificar Efectos Adversos (EEAA) a medicamentos. Uno de los problemas importantes a los que se enfrenta la farmacología es el de la detección de EEAA, algo que produce grandes pérdidas personales y económicas. La detección de estos EEAA es un caso especial de diagnósticos CIE-10 que cuenta además con la particularidad de que, en muchas ocasiones, estos efectos no son codificados adecuadamente, ya que el personal médico no siempre diagnostica estos EEAA al no ser en muchos casos la causa principal de tratamiento, y dada la premura de tiempo en la que se mueve el personal que realiza la codificación. Por ello, el desarrollo de herramientas automáticas capaces de identi-

[3] http://nlp.uned.es/

[4] http://http://www.talp.upc.edu/

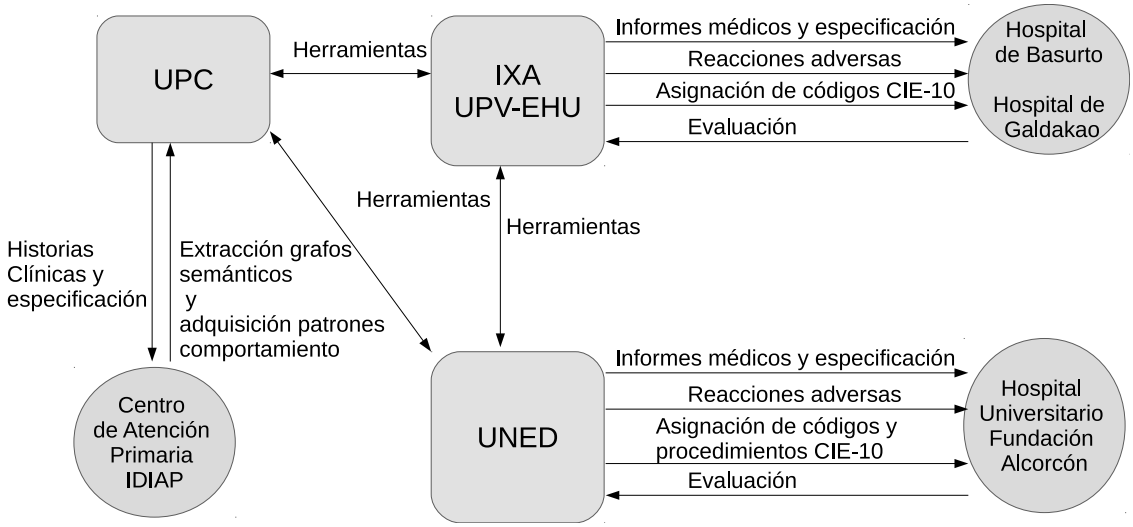

Figura 1: Esquema de colaboración entre los grupos

ficar este tipo de relaciones entre medicamentos y enfermedades puede suponer un importante avance.

- Aunque la lengua en la que se ha desarrollado una mayor cantidad de recursos es el español, este proyecto realizará un esfuerzo en el desarrollo de recursos y herramientas de procesamiento médico para el catalán y el vasco, de manera que se avance en el tratamiento multilingüe de los informes médicos.
- Desarrollar una metodología para la adquisición de grafos semánticos relativos a historias clínicas. Las historias clínicas de cada paciente contienen información textual sobre la evolución clínica del paciente y el análisis de dicha información puede ser de interés relevante para el desarrollo de futuras actuaciones clínicas. Por ello, el desarrollo de una metodología capaz de obtener grafos semánticos donde esa información se representa en formato estructurado, y de adquirir patrones de comportamiento a partir de ellos, puede resultar de gran interés para la comunidad médica en asistencia primaria.

3.1 Casos de uso

Presentamos tres casos de uso específicos de interés para las instituciones médicas que colaboran en el proyecto:

1. Codificación automática de informes médicos con códigos CIE-10.
2. Detección de reacciones adversas a medicamentos.
3. Detección de relaciones entre conceptos que permitan descubrir nuevo conocimiento médico.

El tipo de relación identificada en el caso 2 será primordial para facilitar y mejorar la solución del caso 1 y ambos, a su vez, se utilizarán en el caso 3 para establecer patrones sobre el historial clínico de un paciente.

La figura 1 muestra la interrelación entre los subproyectos, entidades colaboradoras y los casos de uso.

Agradecimientos

Esta contribución ha sido subvencionada por el MINECO (TIN2016-77820-C3-1-R, TIN2016-77820-C3-2-R, TIN2016-77820-C3-3-R y AEI/FEDER, UE.)

Bibliografía

Díaz, A., K. Gojenola, L. Araujo, y R. Martínez. 2015. Extracción de relaciones entre conceptos médicos en fuentes de información heterogéneas (extrecm). *Procesamiento del Lenguaje Natural*, 55:157–160.

World Health Organization. 2009. International statistical classification of diseases and related health problems. Geneva: World Health Organization.