DOTT-HEALTH: Desarrollo de Tecnología aplicada a textos para el soporte de diagnosis, prevención y gestión de instituciones de SALUD

DOTT-HEALTH: Development Of Text-based Technology to support diagnosis, prevention and HEALTH institutions management

Lourdes Araujo,[1] **Juan Martínez-Romo**[1] **Jordi Turmo**[2]
Lluis Padró[2] **Arantza Casillas**[3] **Koldo Gojenola**[3]
[1]UNED. NLP & IR group. IMIENS. C/ Juan del Rosal, 16 28040 Madrid
[2]UPC. TALP Research Center, IDEAI Research Center
C/ Jordi Girona, 1-3 08034 Barcelona
[3]UPV/EHU. HiTZ Center for Language Technologies,
P. M. Lardizabal, 1, 20018 San Sebastián
[1]{lurdes,juaner}@lsi.uned.es [2]{turmo,padro}@cs.upc.edu [3]{arantza.casillas,koldo.gojenola}@ehu.eus

Resumen: La combinación de datos y pautas dirigidas a pacientes individuales se engloba en los Sistemas de Apoyo a la Decisión Clínica. La adopción del Informe Clínico Electrónico de forma sistemática por parte de los sistemas de salud da lugar a una recopilación masiva de datos clínicos que los profesionales no pueden procesar, dada la limitación humana para manejar una gran cantidad de información. Esto, junto con el aumento de la capacidad de procesamiento de las máquinas, conduce a un escenario en el que el análisis automático de los Informes Clínicos Electrónicos se vuelve esencial para determinar patrones, prevenir errores, mejorar la calidad, reducir costos y ahorrar tiempo a los servicios de salud. Esta propuesta aborda dos desafíos principales: el desarrollo de tecnologías para el apoyo al diagnóstico clínico y a la prevención, y la creación de tecnologías de ayuda a la gestión de los servicios médicos. Teniendo todo esto en mente, el proyecto se enfocará en desarrollar herramientas que supongan un avance de la tecnología en los sistemas de apoyo para la toma de decisiones médicas.
Palabras clave: Sistemas de apoyo a la decisión clínica, minería de datos, extracción de información, fenotipado de pacientes, grafos semánticos, aprendizaje profundo.

Abstract: The combination of individual patient data and guidelines is conceptualized as clinical decision support systems. The increase in the adoption of Electronic Health Records (EHR) by healthcare systems results in a collection of massive healthcare data that practitioners, having a limited capability to deal with a big amount of information, are unable to process. This, together with the increase of machine processing capabilities, leads to a scenario where automatic analysis of Electronic Health Records becomes essential to ascertain patterns, to prevent errors, improve quality, reduce costs and save time to the Health Services. This proposal addresses two main challenges: Development of technologies to support the clinical diagnosis and prevention, and to support the management of medical services.
Keywords: Clinical decision support systems, data mining, information extraction, patient phenotype, semantic graphs, deep learning.

1 Descripción general

El proyecto DOTT-HEALTH[1] es un proyecto financiado por el Ministerio de Ciencia e Innovación en la convocatoria 2019 de Proyectos I+D+i, dentro del Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad. DOTT-HEALTH. Es un proyecto coordinado entre la Universidad del País Vasco, la Universidad Nacional de Educación a Distancia y la Universidad Politécnica de Cataluña, que en algunos objetivos es continuación de proyectos anteriores del consorcio relacionados con aplicaciones del procesamiento del lenguaje en el dominio de la salud, como PROSAMED (Díaz de Ilarraza et al., 2017) y EXTRECM (Díaz de Ilarraza et al., 2015). En este nuevo proyecto, la investigación y los distintos casos de uso se relacionan con el desarrollo de sistemas de ayuda a la toma de decisiones en distintos ámbitos de salud: psiquiatría, multimorbilidad o fenotipado. En la figura 1 se da una visión general del proyecto.

El soporte a la toma de decisiones clínicas (CDS) tiene como objetivo ayudar a los médicos, al personal, a los pacientes y a los proveedores de atención sanitaria a mejorar la salud y la atención sanitaria proporcionando conocimientos e información filtrada de forma inteligente.

En general, un CDS puede conducir a mejoras significativas en los servicios de salud, afectando a la seguridad, la eficiencia y la eficacia de la atención sanitaria. La evolución y el aumento del uso de los sistemas de CDS en la práctica son inevitables (Middleton, Sittig, y Wright, 2016), dada la explosión de información biomédica y la presión para mejorar la calidad y reducir los costes de la atención basada en el valor. Sin embargo, es necesario seguir trabajando en la estandarización de los métodos de representación de datos, en la construcción de sistemas transparentes y en el intercambio de datos y conocimientos de los pacientes.

A pesar de la extensa investigación en el tema, las aplicaciones de CDS no han alcanzado una amplia aceptación y utilización en el ámbito de atención sanitaria debido, entre otras razones (Wagholikar, Sundararajan, y Deshpande, 2012), a los desafíos que plantea la necesidad de utilizar técnicas de Procesamiento de Lenguaje Natural PLN, ya que la mayoría de la información de los pacientes no está estructurada sino que se encuentra en forma de texto libre. Con los avances en este ámbito, el texto libre puede proporcionar información útil que se puede integrar con otras procedentes, por ejemplo, de pruebas de laboratorio, mejorando así la precisión de las decisiones de las aplicaciones de CDS.

En este proyecto nos proponemos avanzar en técnicas de PLN que permitirán mejorar los CDS en el ámbito sanitario. Como valor adicional, esto llevará a la mejora de las técnicas de procesamiento de informes clínicos electrónicos (ICE) que registran un tipo de texto con características muy específicas. En concreto, entre las técnicas que exploraremos se encuentran la anonimización, la negación y especulación, la desambiguación de acrónimos, cuya presencia en los informes médicos es extremadamente común, la identificación de expresiones temporales que permiten la creación de líneas de tiempo de eventos relacionados con un paciente, y la exploración del análisis de estados emocionales como fuente de información adicional sobre el paciente.

También se avanzará en la mejora de los modelos de representación de la información para el dominio biomédico (por ejemplo en la combinación de *embeddings*, grafos, reglas de asociación, etc.), así como en el enriquecimiento de ontologías médicas con variantes terminológicas de conceptos y su búsqueda aproximada, que permitirá mejorar en el resto de objetivos. Las herramientas y la tecnología desarrolladas conjuntamente por los grupos serán aplicadas y evaluadas sobre algunos casos de uso de alta relevancia médica y, por tanto, social.

Entre los casos de uso considerados está la ayuda a la predicción de casos de riesgo en psiquiatría (suicidio, autolesiones, consumo de sustancias, aislamiento social, abandono del tratamiento, etc.), que explorará técnicas de detección de estados de ánimo que complementarán la información relacionada con las entidades médicas mencionadas en el ICE.

Otro caso de uso que se considerará es la predicción de riesgo de multimorbilidad. Dado que la Organización Mundial de la Salud sostiene que la multimorbilidad esta mostrando una mayor prevalencia en las últimas décadas y en países de toda índole, la capacidad de evaluar el riesgo de multimorbilidad no solo reduciría la proporción de ciudadanos afec-

[1] http://www.ixa.eus/node/13110?language=en

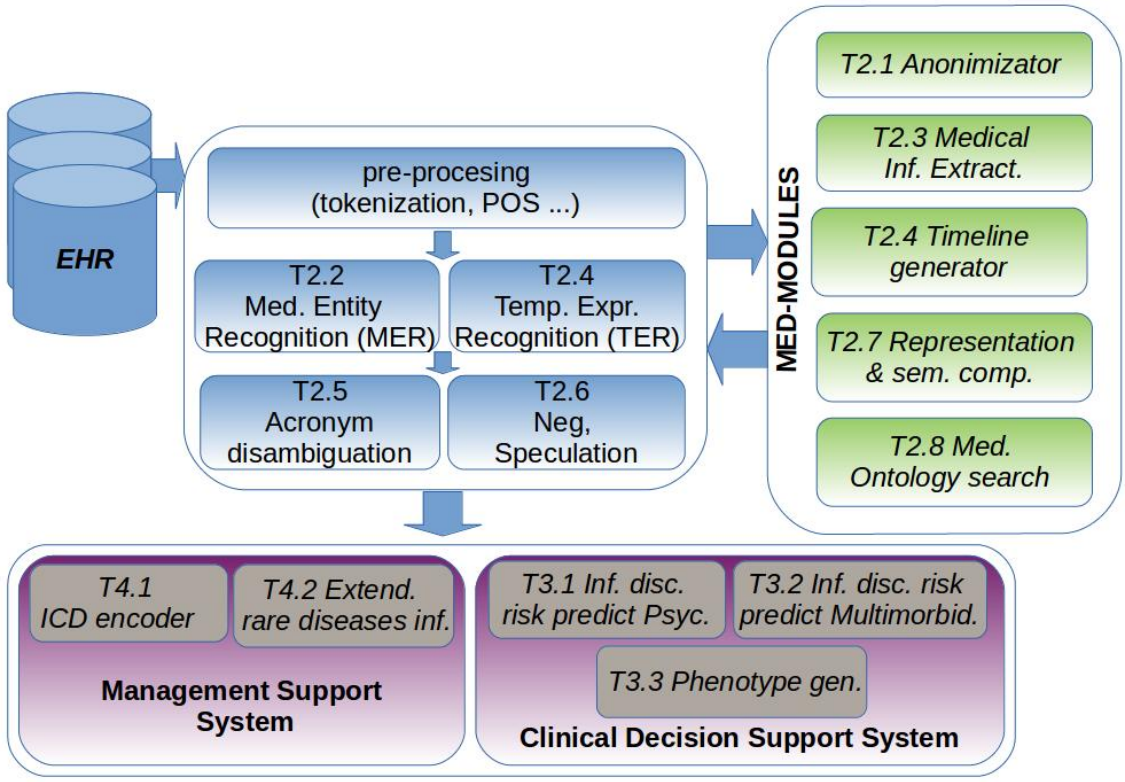

Figura 1: Descripción general del proyecto.

tados, sino que también mejoraría su calidad de vida y reduciría costes sanitarios.

También se abordará el fenotipado de los pacientes, ya que puede ayudar a identificar correctamente la cohorte de pacientes y a identificar mejor el contexto clínico. Esto puede suponer avances fundamentales en la gestión médica, la prevención y el diagnóstico identificando cohortes de pacientes y desarrollando medidas de similitud entre pacientes, lo que puede aportar nuevos conocimientos importantes.

También trataremos la tecnología enfocada a la gestión hospitalaria. Se abordará el problema de la identificación de casos de enfermedades raras (ER) que no han sido identificados por su nombre genérico, mejorando así la exhaustividad de su registro. Además, buscaremos correlaciones con malformaciones congénitas, una indicación que puede ser de gran ayuda para identificar casos de ER no mencionados explícitamente en los registros. También seguiremos avanzando en la clasificación y recomendación de los códigos de la CIE-10, un problema complejo que se empezó a investigar en el proyecto anterior del consorcio.

2 Grupos Involucrados

DOTT-HEALTH consta de tres subproyectos:

- PAT-MED: PATient characterization and MEDical document management through text-based technology.
- INDICA-MED: INformation DIscovery and CAtegorization based on language processing for the MEDical domain.
- TADIA-MED: Medical Text Analysis for Disease Prediction Assistance.

El proyecto, que tiene naturaleza multidisciplicar, cuenta con la colaboración de tres grupos de investigación en PLN y varias instituciones relacionadas con salud, con las que se desarrollarán los casos de uso de cada subproyecto.

Los grupos involucrados en los subproyectos son:

- Grupo IXA[2] de la Universidad del País

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

Vasco Vasco UPV/EHU. Lleva trabajando cerca de treinta años en Procesamiento de Lenguaje Natural (PLN) y Lingüística Computacional en general. Desde hace diez años viene desarrollando una línea de investigación orientada al trabajo con textos médicos.

- Grupo NLP&IR[3] de la UNED. Cuenta con una larga experiencia en Acceso Inteligente a la Informacióon y Adquisición y Representación de Conocimiento. En particular tiene diversas 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.
- Hospitales públicos Universitarios Clínico San Carlos y Fundación Universitaria Fundación Alcorcón (HUFA) de la Comunidad de Madrid. Estos hospitales, junto con la Consejería de Sanidad de la Comunidad de Madrid colaboran en el subproyecto INDICA-MED del grupo UNED.
- Fundación IDIAP JGol, integrada en el grupo TALP. IDIAP desarrolla y gestiona la investigación de la Atención Primaria principalmente en Cataluña, facilitando la participación de investigadores de distintos sectores.

3 Objetivos

El proyecto plantea lograr estos objetivos:

- Desarrollo de tecnología para el soporte al diagnóstico y prevención: extracción de información médica (entidades médicas, negación e incertidumbre) y descubrimiento de patrones para prevención y diagnosis. Identificación de relaciones relevantes entre conceptos médicos y extracción de patrones temporales en registros históricos de pacientes.
- Desarrollo de tecnología para CDS y soporte a la gestión de instituciones de salud: clasificación de acuerdo a sistemas de codificación médica (p.e., CIE.10), extensión de información sobre enfermedades raras, predicción de multimorbilidad y fenotipado.

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

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

Agradecimientos

Este trabajo ha sido financiado por el proyecto DOTT-HEALTH (MCI/AEI/FEDER, UE) con referencias PID2019-106942RB-C31, PID2019-106942RB-C32, PID2019-106942RB-C33.

Bibliografía

Díaz de Ilarraza, 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(0):157–160.

Díaz de Ilarraza, A., K. Gojenola, R. Martínez, V. Fresno, J. Turmo, y L. Padró. 2017. Procesamiento semántico textual avanzado para la detección de diagnósticos, procedimientos, otros conceptos y sus relaciones en informes medicos (PROSA-MED). *Proces. del Leng. Natural*, 59:133–136.

Middleton, B., D. Sittig, y A. Wright. 2016. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearbook of medical informatics*, Suppl 1:S103–16, 08/2016.

Wagholikar, K. B., V. Sundararajan, y A. W. Deshpande. 2012. Modeling paradigms for medical diagnostic decision support: a survey and future directions. *Journal of medical systems*, 36(5):3029—3049, October.