Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 65-71 recibido 04-04-2018 revisado 02-05-2018 aceptado 04-05-2018

Plataforma para la extracción automática y codificación de conceptos dentro del ámbito de la Oncohematología (Proyecto COCO)

Platform for the automatic extraction and coding of concepts within the scope of Oncohematology (COCO Project)

Silvia Sánchez Seda[1], Francisco de Paula Pérez León[1], Jesús Moreno Conde[1], María C. Gutiérrez Ruiz[1], Jesús Martín Sánchez[2], Guillermo Rodríguez[2], José Antonio Pérez Simón[2], Carlos L. Parra Calderón[1]

[1] Grupo de Innovación Tecnológica del Hospital Universitario Virgen del Rocío de Sevilla
[2] Unidad de Gestión Clínica de Hematología del Hospital Universitario Virgen del Rocío de Sevilla
(jesus.moreno.conde.exts@juntadeandalucia.es)

Resumen: El proyecto COCO tiene como objetivo diseñar, desarrollar y validar un sistema de extracción de conocimiento que, a partir de los textos de la Historia de Salud Electrónica, codifique automáticamente los diagnósticos de Oncohematología mediante Tecnologías del Lenguaje basada en un estándar de pipeline interoperable. La necesidad de normalizar el conocimiento de la Historia Clínica constituye un gran desafío. Puesto que la CIE-10 presenta limitaciones para representar esta información, se desarrolló la norma CIE-O-3, para dar soporte a este tipo de patologías. Se propone desarrollar el primer pipeline de Procesamiento del Lenguaje Natural de componentes interoperables, así como, el primer codificador automático CIE-O-3 y CIE-10. Nuestro sistema servirá de apoyo a la decisión, investigación y gestión clínica en este campo.
Palabras clave: Recuperación de conocimiento clínico, Codificación Automática, Procesamiento del Lenguaje Natural, Oncohematología, CIE-10, CIE-O-3

Abstract: The COCO project aims to design, develop and validate a knowledge extraction system that, based on the texts of the Electronic Health Record, automatically codes Oncohematology diagnostics using Language Technologies based on an interoperable pipeline standard. The need to standardize knowledge of the Electronic Health Record is a major challenge. Since ICD-10 has limitations in representing this information, ICD-O-3 was developed to support this type of pathology. It is proposed to develop the first Natural Language Processing pipeline of interoperable components, as well as the first ICD-O-3 and ICD-10 automatic encoder. Our system will support clinical decision making, research and management in this field.
Keywords: Recovery of clinical knowledge, Automatic Coding, Natural Language Processing, Oncohematology, ICD-10, ICD-O-3

1 Introducción

El tratamiento de la información no estructurada en el sector sanitario es un gran desafío, a la par que, muy prometedor. El uso de información estructurada en los hospitales es escaso, y la mayor fuente de información textual son los informes escritos por los médicos después de una consulta o un procedimiento, los cuales se expresan en texto libre. Estos informes narrativos (e.g., informes de consulta, patología, radiología, etc.) contienen información valiosa para el diagnóstico que puede ayudar al pronóstico y a predecir el comportamiento biológico de las enfermedades (Mohanty, 2007). En este sentido, varios estudios han señalado la viabilidad del uso de técnicas de Procesamiento del Lenguaje Natural (PLN) para la estructuración de informes de texto libre (Spasić, 2014) (Buckley, 2012) (Warner, 2011)

ISSN 1135-5948. DOI 10.26342/2018-61-7

(Martinez, 2011). Evaluaciones de estos sistemas sugieren que pueden proporcionar datos precisos sobre la prestación de servicios y el estado clínico del paciente (Chan, 2010). En particular, las técnicas de PLN pueden ser relevantes para la investigación y soporte a la decisión clínica, al permitir la recuperación de información, así como la respuesta a preguntas en el contexto de una multitud de datos disponibles. Todo ello, eliminando el uso intensivo de recursos para la revisión manual (Tange, 1998).

La gran mayoría de los trabajos citados se han centrado en el idioma inglés, no así en el español, que es el segundo idioma más hablado del mundo hoy en día. En el ámbito de PLN en español aplicado a medicina se encuentran solamente algunos trabajos (Menasalvas, 2016) (Costumero, 2014) (Oronoz, 2013) (Medrano, 2018), mientras que el número especializado en el área de la codificación automática de diagnósticos es más reducido. Existen aplicaciones comerciales como Savanamed (Espinosa, 2016) y MeaningCloud que realizan codificaciones automáticas en el dominio de la medicina en español, pero ninguna de ellas realiza una codificación automática en un dominio más especializado como es la Oncohematología, haciendo uso de la CIE-O-3 ya que es un dominio para el que la CIE-9 o CIE-10 presentan limitaciones.

La CIE-10 es la décima edición de la Clasificación internacional de enfermedades, que se corresponde con la versión española de la ICD (International Statistical Classification of Diseases and Related Health Problems), y determina la clasificación y codificación de las enfermedades y una amplia variedad de signos, síntomas, hallazgos anormales, denuncias, circunstancias sociales y causas externas de daños y/o enfermedad. Sin embargo, esta codificación no facilita códigos de morfología en el índice alfabético de enfermedades ni dispone de un apéndice de morfologías, por lo que, para la adecuada codificación de la morfología de los tumores, se deberá acudir a la clasificación CIE-O. La CIE-O es una clasificación dual, con sistemas de codificación tanto para la topografía como para la morfología. El código topográfico de la CIE-O describe el sitio de origen de las neoplasias. El código de morfología describe el tipo de células del tumor y su actividad biológica; en otras palabras, las características del tumor mismo. En la actualidad, se utiliza la tercera edición de la CIE-O (CIE-O-3).

En la literatura, tan sólo encontramos dos trabajos orientados a codificar de forma automática la Historia de Salud Electrónica **(**HSE) en base a CIE-O (Jouhet, 2012) (Kavuluru, 2013). En (Jouhet, 2012) se clasifican informes de anatomía patológica en base CIE-O-3 utilizando para ellos técnicas de Aprendizaje Automático. En estas técnicas, como algoritmos de clasificación se utilizaron Naïve Bayes y SVM, los cuáles fueron evaluados sobre 5.121 informes de anatomía patológica elaborados por 35 profesionales médicos. Para medir la eficacia del sistema desarrollado, se tuvo en cuenta la capacidad del mismo para atribuir correctamente un código preciso de la CIE-O-3, tanto para los ejes topográficos como morfológicos. El sistema obtuvo una medida-F de 71,5% para topografía y 85,4% para morfología, estos resultados, sugieren que los informes de anatomía patológica podrían ser útiles como fuente de datos para sistemas automatizados con el fin de identificar y notificar nuevos casos de cáncer. Además, se hace evidente la necesidad de trabajos futuros que incluyan técnicas de PLN así como la incorporación de otros tipos de documentos médicos.

Por su parte, en (Kavuluru, 2013) se codificó automáticamente el diagnóstico en base a CIE-O-3 para un total de 56.426 informes de anatomía patológica asociados con casos de cáncer almacenados en el Registro de Cáncer de Kentucky, y procedentes de 35 laboratorios diferentes. En este documento, sólo un código CIE-O-3 se vinculó con cada informe de anatomía patológica, y los métodos de Aprendizaje Automático implementados obtuvieron una medida-F de 90%. Los autores del artículo manifiestan la necesidad de seguir investigando en esta línea con objeto de perfeccionar el sistema y extenderlo a otros dominios. Estos trabajos están centrados en el dominio del inglés y según nuestro conocimiento, no existe ningún trabajo previo que aplique estas técnicas en historias clínicas en español, de ahí la relevancia y el impacto de la innovación propuesta. Esta propuesta supondría desarrollar, en el dominio del español, el primer pipeline de PLN de componentes interoperables basados en el estándar UIMA en castellano para extraer conocimiento de la HSE así como también implementar el primer codificador automático en base a CIE-O-3 y CIE-10, sirviendo de experiencia temprana para el desarrollo de procesos de Compra Pública de Innovación

sobre TL en Sanidad tanto por parte de la Administración Pública Andaluza como de otras Comunidades Autónomas en el Marco del Plan de Impulso de Tecnologías del Lenguaje del MINETAD.

2 Objetivo

El proyecto COCO tiene como objetivo diseñar, desarrollar y validar un sistema de extracción de conocimiento a partir de la información contenida en la HSE. Dicho sistema, mediante técnicas de Aprendizaje Automático, se centrará en la codificación automática de diagnósticos en el dominio de Oncohematología, sirviendo como soporte a la toma de decisiones clínicas, además de proporcionar grandes ventajas tanto a la investigación como a la gestión clínica. Para ello este proyecto persigue:

- Desarrollar un pipeline de componentes interoperables que apliquen técnicas de PLN adaptadas al dominio clínico en español capaz de procesar el texto libre para estructurar la información contenida en la HSE.
- Desarrollar mediante técnicas de Aprendizaje Automático, nuevos modelos de clasificación que permitan la codificación automática del diagnóstico, en base a CIE-0, asociado a la HSE de los pacientes de Oncohematología, mapeando dicha codificación con la clasificación CIE-10 utilizada en el Sistema Sanitario Público de Andalucía (SSPA) a la hora de codificar los diagnósticos asociados a los pacientes.

3 Material y Métodos

Este proyecto propone un estudio compuesto por 6 fases principales que se desarrollarán en 24 meses:

Fase 1. Identificación de los sujetos de estudio y análisis preliminar de las HSE.

- Identificación de los sujetos de estudio.
- Análisis preliminar de la HSE.
- Proceso de extracción, transformación y carga (ETL) de las HSE.

Fase 2. Desarrollo y ejecución de componentes interoperables que apliquen técnicas de Procesamiento del Lenguaje Natural (PLN) para la extracción de conocimiento de la HSE.

- Creación de un corpus ad-hoc para la extracción del conocimiento de la HSE.
- Desarrollo de módulos interoperables que permitan realizar el workflow completo de PLN para extraer el conocimiento de las HSE.
- Evaluación de la precisión general del sistema para verificar la adecuación de los procesos y herramientas generados.

Fase 3. Estudio e identificación de la información estructurada y análisis combinado del conocimiento generado.

- Identificación de información estructurada.
- Integración de datos estructurados identificados en esquema común.

Fase 4. Creación, análisis y despliegue de modelos de clasificación para la codificación automática de diagnósticos.

- Análisis descriptivo de datos recopilados.
- Creación de modelos para el Aprendizaje Automático.
- Mapeo entre CIE-O-3 y CIE-10.
- Evaluación de los modelos generados.
- Despliegue de los modelos más prometedores.

Fase 5. Integración de la información en un data-warehouse orientado a la investigación clínico y traslacional y gestión clínica.

- Diseño del Data Warehouse.
- Implementación del Data Warehouse.

Fase 6. Validación e implementación del sistema en el entorno de la Unidad de Gestión Clínica (UGC) de Hematología del Hospital Universitario Virgen del Rocío de Sevilla (HUVR).

- Despliegue del sistema.
- Evaluación del sistema.

La población principal de sujetos de estudio estará compuesta por datos anonimizados de pacientes que acudan al Servicio de Oncohematología del HUVR desde el año 2013 hasta el fin de 2016. Se incluirán los pacientes con algún episodio en el Servicio de Oncohematología del HUVR y se excluirán aquellos con falta de datos relevantes.

Actualmente, el proyecto COCO se encuentra en las primeras 2 fases del proyecto donde se centran los trabajos en la identificación y creación del corpus de trabajo y en el diseño y desarrollo de la arquitectura de componentes del sistema.

3.1 Arquitectura

El diseño del sistema COCO presenta una arquitectura similar a la que se presenta a continuación:

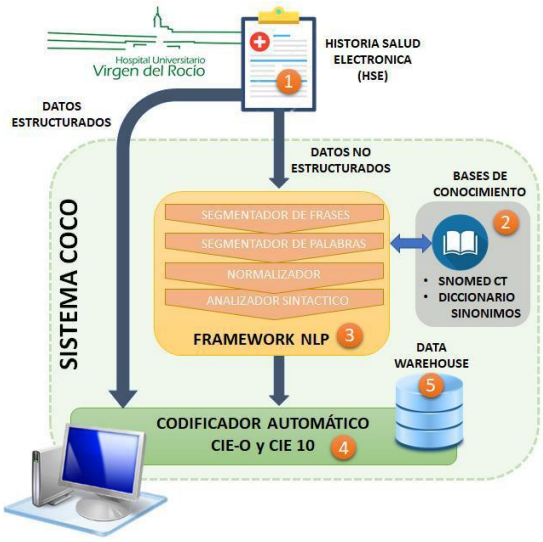

Figura 1: Representación de la arquitectura de componentes del sistema COCO

(1) HSE: informes clínicos que contienen la información procedente de los sistemas de Historia Clínica Electrónica del Hospital. Esta información podrá ser información de texto libre no estructurada que pasará por los algoritmos de procesamiento de lenguaje natural, así como códigos e información estructurada que permitirá mejorar la codificación automática de los códigos de diagnósticos de los pacientes

(2) Bases de conocimiento: servicio interno que contiene información sobre las terminologías de referencia dentro del ámbito de la oncohematología. Además, nos servirá para identificar los conceptos relevantes en el texto libre extraído de la HSE, así como, información relacionada con los mismos (Ej. Diccionarios de sinónimos, acrónimos, estructuras gramaticales, etc.)

(3) Framework NLP: Servicio Web basado en Apache UIMA y dividido en varios módulos encargados de realizar una tarea concreta en el procesamiento del texto (Ej. segmentación, normalización, tokenización, análisis, etc.). Para la realización de este framework se está valorando sustituir Apache UIMA por Python (mediante la librería NLTK) como lenguaje de referencia para el desarrollo de tareas de procesamiento de lenguaje natural. A través de este framework se implementará una interfaz de servicios para la ejecución de tareas de PLN sobre el texto libre contenido en la HSE de los pacientes, garantizando la escalabilidad y reutilización del sistema, a través de un Pipeline de trabajo segmentado.

(4) Codificador automático: algoritmos basados en técnicas avanzadas de Data Mining y aprendizaje automático para la identificación de conceptos CIE-O a partir de los conceptos identificados en el Framework NLP y los datos estructurados recogidos en la HSE de los pacientes.

(**5) DataWareHouse**: sistema de almacenamiento y análisis de la información clínica de los pacientes obtenida de implementar los procesos de PLN. Este sistema ofrecerá una interfaz de trabajo sobre la que implementar consultas guiadas sobre los datos registrados con el fin de facilitar la explotación de la información contenida en el sistema.

3.2 Corpus

El corpus trabajo generado dentro del marco del proyecto consiste en informes clínicos procedentes del Hospital Virgen del Rocío de Sevilla. De todos ellos, se distinguen informes únicos de alta, informes de consulta y hojas de evolución, todos ellos escritos en texto libre. Los informes fueron cogidos aleatoriamente entre informes del año 2013 al 2016 y fueron completamente anonimizados. El corpus completo fue anotado por dos expertos del dominio, siguiendo las guías de anotación que desarrollamos previamente. Algunos documentos del corpus no fueron incluidos en el corpus final, ya que sirvieron de entrenamiento para los anotadores. La herramienta de anotación usada fue la que ofrece APACHE UIMA y la tarea de anotación consistió en identificar los eventos clínicos relacionados con alguno de los 160 conceptos CIE-O-3 que entraban dentro del alcance del

proyecto, las partículas que expresan negación en la frase y las palabras que están dentro del alcance de cada partícula de negación.

3.3 Flujo de procesamiento NLP

La tecnología utilizada para el procesamiento del lenguaje es Apache UIMA, el único estándar reconocido por OASIS (Organización para el Avance de Estándares de Información Estructurada). El desarrollo bajo el estándar Apache UIMA permite la interoperabilidad y los distintos componentes de PLN podrían ser implementados por distintas organizaciones. También se plantea el uso de Python y la librería NLTK, ya que es un lenguaje de fácil uso y multiplataforma.

Figura 2: Representación del pipeline del procesamiento del texto

Se entrenarán 160 clasificadores binarios, uno por cada código CIE-O-3.
Se realizará un proceso de evaluación para el sistema desarrollado de forma que se considera que un registro está correctamente codificado si el clasificador del código que le corresponde lo clasifica como positivo y los clasificadores de los códigos que no le corresponden lo clasifican como negativo. Para este proceso de evaluación se usará un subconjunto de documentos para la fase de entrenamiento del sistema y el resto de documentos para la fase de evaluación de los algoritmos. En la fase de evaluación se analizarán parámetros de calidad como la precisión, la exhaustividad y el valor-F para cada clasificador.

4 Resultados

Como resultado esperado, se prevé desarrollar un nuevo producto que servirá de apoyo tanto a procesos de soporte a la decisión clínica como a tareas de investigación y gestión clínica dentro de la UGC de Hematología del Hospital Universitario Virgen del Rocío. Además, se espera que este sistema se integre en la práctica clínica habitual de la Unidad, pudiendo replicarse en otros dominios y centros del SSPA. Este hecho favorecería importantes mejoras al SSPA, entre las que se destacan:

- Reducción de costes: Se produciría una disminución de los recursos necesarios para llevar a cabo la codificación de la HSE (personas involucradas en el proceso/mes). Además, se reducirían también los costes asociados a la formación de codificadores.
- Disminución de errores. Se reduciría el número de errores que los anotadores humanos cometen al tener que trabajar con miles de posibles códigos al asignar las etiquetas de la CIE-10 a un documento.
- Automatización de procesos de codificación. La automatización del proceso de asignación de códigos CIE-10 a la HSE aumentaría el número de historias clínicas que se codifican.
- Mejora en la capacidad de análisis epidemiológico o estadístico: Se podrían tener datos sobre la incidencia y prevalencia de las distintas patologías en una población dada. Además, se mejoraría el conocimiento profundo y actualizado de la información gestionada por la UGC de Hematología para la dirección de políticas públicas.
- Mejora en la gestión de la Unidad: Facilitaría las tareas de gestión en la UGC de Hematología para el cálculo de sus indicadores anuales. Además, el sistema de gestión sanitaria obtendría de origen la información clínica con los diagnósticos codificados, con todas las ventajas que esto supone.
- Aumento de la investigación: Facilitaría la identificación de cohortes de pacientes para el desarrollo de ensayos clínicos en el HUVR. La estandarización de la información en texto libre contenida en la HSE facilitaría su explotación por

otros servicios del SSPA. Esta información podría ser compartida con otros sistemas sanitarios del SSPA lo que permitiría avanzar en la investigación de las hemopatías malignas, patologías graves con poca frecuencia.

- Mejora de la riqueza semántica de los sistemas de Historia Clínica. La UGC de Hematología contaría con la información asociada a las hemopatías malignas normalizada de forma automática a nivel internacional según el estándar CIE-O-3.
- Promoción del uso de herramientas que faciliten el desarrollo de procesos de PLN en el SSPA. El desarrollo de un pipeline de PLN de componentes interoperables permitiría ejecutar las distintas fases (segmentador de frases, segmentador de palabras, etiquetador gramatical, etc.) de forma independiente en función de las necesidades de las tareas a resolver. Al estar desarrollado bajo el estándar UIMA, los distintos componentes de PLN interoperables podrían ser implementados por distintas organizaciones.
- Proveer al SSPA de recursos lingüísticos reutilizables dentro de la política de Reutilización de la Información del Sector Público (RISP).
- Impulsar el desarrollo de herramientas de soporte a la decisión para la personalización de tratamientos en base a pacientes con diagnósticos similares, nuevos servicios de información que apliquen algoritmos predictivos, etc.

5 Discusión

Entre los trabajos previos, centrados en el ámbito del español, solamente se encuentran algunas aplicaciones comerciales incipientes sin información detallada como Savanamed (Espinosa-Anke, 2016) o MeaningCloud que hacen uso de PLN, del Aprendizaje Automático y otras técnicas para beneficiar al sector de la salud y extraer conocimiento de la HSE. En cuanto a los trabajos orientados a codificar de forma automática la HSE en base a CIE-O-3, no existe ningún trabajo previo para el español. Por tanto, el sistema COCO presenta una clara novedad en el SSPA pues no hay ninguna iniciativa parecida que resuelva el problema planteado. Esto, supone una gran mejora, no sólo desde el punto de vista económico sino desde el punto de vista de la investigación, la práctica clínica y traslacional, la gestión, etc., y servirá de experiencia temprana para el desarrollo de procesos de Compra Pública de Innovación sobre TL en Sanidad, tanto por parte de la Administración Pública Andaluza como de otras Comunidades Autónomas en el Marco del Plan del MINETAD.

Agradecimientos

Esta investigación ha sido financiada en parte por la Plataforma de Innovación en Tecnologías Médicas y Salud (Plataforma ITEMAS, PT13/0006/0036) financiado por el Instituto de Salud Carlos III y por el proyecto COCO (PIN-0121-2017) financiado por la consejería de Salud de la Junta de Andalucía ambos cofinanciados a través de los Fondos Europeos de Desarrollo Regional (FEDER).

Bibliografía

Mohanty, S. K., Piccoli, A. L., Devine, L. J., Patel, A. A., William, G. C., Winters, S. B., y Parwani, A. V. 2007. Synoptic tool for reporting of hematological and lymphoid neoplasms based on World Health Organization classification and College of American Pathologists checklist. *Bmc Cancer*, 7(1), 144.

Spasić, I., Livsey, J., Keane, J. A., y Nenadić, G. 2014. Text mining of cancer-related information: review of current status and future directions. *International journal of medical informatics*, 83(9), 605-623.

Buckley, J. M., Coopey, S. B., Sharko, J., Polubriaginof, F., Drohan, B., Belli, A. K., y Specht, M. C. 2012. The feasibility of using natural language processing to extract clinical information from breast pathology reports. *Journal of pathology informatics*, 3.

Warner, J. L., Anick, P., Hong, P., y Xue, N. 2011. Natural language processing and the oncologic history: is there a match? *Journal of oncology practice*, 7(4), e15-e19.

Martinez, D., y Li, Y. 2011. Information extraction from pathology reports in a hospital setting. In *Proceedings of the 20th ACM international conference on*

Information and knowledge management, pages 1877-1882. ACM.

Chan, K. S., Fowles, J. B., y Weiner, J. P. 2010. Electronic health records and the reliability and validity of quality measures: a review of the literature. *Medical Care Research and Review*, 67(5):503-527.

Tange, H. J., Schouten, H. C., Kester, A. D., & Hasman, A. 1998. The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *Journal of the American Medical Informatics Association*, 5(6):571-582.

Menasalvas, E., Rodriguez-Gonzalez, A., Costumero, R., Ambit, H., y Gonzalo, C. 2016. Clinical Narrative Analytics Challenges. In *International Joint Conference on Rough Sets*, pages 23-32. Springer, Cham.

Costumero, R., García-Pedrero, Á., Gonzalo-Martín, C., Menasalvas, E., y Millan, S. 2014. Text analysis and information extraction from Spanish written documents. In *International Conference on Brain Informatics and Health*, pages 188-197. Springer, Cham.

Oronoz, M., Casillas, A., Gojenola, K., y Perez, A. 2013. Automatic annotation of medical records in Spanish with disease, drug and substance names. In *Iberoamerican Congress on Pattern Recognition*, pages 536-543. Springer, Berlin, Heidelberg.

Medrano, I. H., Guijarro, J. T., Belda, C., Ureña, A., Salcedo, I., Espinosa-Anke, L., y Saggion, H. 2018. Savana: Re-using Electronic Health Records with Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4 (Special Issue on Big Data and e-Health).

Espinosa-Anke, L., Tello, J., Pardo, A., Medrano, I., Ureña, A., Salcedo, I., y Saggion, H. 2016. Savana: A Global Information Extraction and Terminology Expansion Framework in the Medical Domain.

Jouhet, V., Defossez, G., Burgun, A., Le Beux, P., Levillain, P., Ingrand, P., y Claveau, V. 2012. Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods of information in medicine*, 51(3):242.

Kavuluru, R., Hands, I., Durbin, E. B., y Witt, L. 2013. Automatic extraction of ICD-O-3 primary sites from cancer pathology reports. *AMIA Summits on Translational Science Proceedings*, 2013, 112.