Reconocimiento de Entidades en Informes Médicos en Español

Named Entity Recognition in Spanish Medical Reports

Pilar López Úbeda
Sinai Group
Universidad de Jaén
Campus Las Lagunillas s/n. E-23071
plubeda@ujaen.es

Resumen: En los últimos años, los servicios sanitarios españoles están almacenando información acerca de sus pacientes de manera electrónica. Estos documentos contienen información relevante sobre enfermedades, alergias, medicación, etc. El propósito general de este trabajo es reconocer automáticamente entidades relacionadas con la medicina dentro de un informe médico escrito en castellano. En la primera fase utilizaremos técnicas del Procesamiento de Lenguaje Natural (PLN) para el tratamiento del texto para más tarde extraer los términos referidos al dominio de la medicina. Estudiaremos diccionarios y ontologías como SNOMED-CT que contribuyan a resolver los problemas de interoperabilidad semántica y de reutilización de conocimiento en los sistemas de información clínicos.
Palabras clave: procesamiento del lenguaje natural, reconocimiento automático de entidades, dominio médico, ontología, terminología médica, informes médicos

Abstract: During the last years, Spanish health services are storing electronic information about their patients. These documents contain relevant information on diseases, allergies, medication, etc. The main goals of this work is to automatically recognize entities related to medicine in a medical report written in Spanish. In the first phase, we will use Natural Language Processing (NLP) techniques for the treatment of text, and later we will extract the terms of the medical domain. We will study dictionaries and ontologies such as SNOMED-CT, which contribute to solve the problems of both semantic interoperability and knowledge reuse in clinical information systems.
Keywords: natural language processing, automatically recognize entities, medical domain, ontology, medical terminology, medical reports

1 Justificación de la investigación propuesta

Con el paso de los años, la necesidad de tener un buen sistema informático de gestión de informes médicos está teniendo gran importancia, pues los datos contenidos en dichos informes son relevantes. Se hace necesario el diseño y desarrollo de nuevas y potentes herramientas de procesamiento de la información que aprovechen los avances de las tecnologías relacionadas con la información para poder acceder y analizar estos datos, todos ellos relacionados con los registros electrónicos de la salud de un paciente.

Por lo tanto, es común hoy día para un especialista registrar los datos del paciente de manera electrónica. Esto incluye información del paciente, medicación, resultados de análisis, diagnósticos, dosis, etc. Manteniendo esa información digitalizada vemos tres grandes ventajas: se reduce el tiempo de trabajo del personal de salud, se mejora la calidad de la atención y se utiliza la información a través de sistemas automáticos como por ejemplo la minería de textos (Ananiadou and McNaught, 2006).

Llamamos minería de textos al proceso de descubrir, a partir de grandes cantidades de

texto, el conocimiento para encontrar información no trivial, desconocida y potencialmente útil en textos no estructurados.

La minería de textos es un área multidisciplinar donde convergen diferentes paradigmas de computación como son la construcción de árboles de decisión, la inducción de reglas, las redes neuronales, el descubrimiento basado en instancias, programación lógica, algoritmos estadísticos, etc. Las principales tareas y métodos de minería de textos, entre otras, son: extracción de la información, generación automática de resúmenes de textos, categorización, agrupamiento, vinculación entre conceptos, visualización de la información y la respuesta automática de preguntas.

El presente trabajo se centra en el reconocimiento de entidades con nombre, *Named Entity Recognition* (NER), uno de los problemas básicos de la minería de texto (Cohen and Hersh, 2005). Por tanto, se desarrollarán técnicas de extracción de información y de procesamiento del lenguaje natural (PLN) para poder gestionar la información contenida en un expediente médico (Hahn, Romacker, and Schulz, 1999). Como producto final, tendremos información que carece de valor por sí sola, pero podemos llegar a intercambiarla y utilizarla de manera eficiente para obtener conocimiento de ella. A dicha habilidad se la conoce como interoperabilidad semántica. Entendemos por interoperabilidad semántica la capacidad que tienen diferentes sistemas informáticos de compartir información y entenderse.

La importancia de intercambiar información en el dominio clínico, proviene en buena medida de nuevos requisitos que los servicios de atención sanitaria deben afrontar en su organización y funcionalidad para poder seguir prestando su servicio de forma efectiva, eficiente y sostenible, como por ejemplo, los cambios demográficos, movilidad de ciudadanos y la equidad en el acceso.

Para garantizar la interoperabilidad semántica entre sistemas, se hace necesario el uso de estándares que permitan el intercambio de datos, así como la utilización de catálogos estandarizados, los cuales unifican los datos empleados en distintas instituciones derivando en el intercambio correcto de información.

A continuación se mencionan diccionarios y ontologías que contribuyen a resolver los problemas antes mencionados:

- CIE-10: Es la Clasificación Internacional de Enfermedades, décima versión correspondiente a la versión en español de la ICD, por sus siglas en inglés: *International Statistical Classification of Diseases and Related Health Problems*. Se ha convertido en una clasificación diagnóstica estándar internacional para todos los propósitos epidemiológicos generales y es adecuada para clasificar enfermedades y otros tipos de problemas de salud.
- MeSH: *Medical Subjects Headings* es el vocabulario controlado que emplea Medline y otras bases de datos biomédicas para procesar la información que se introduce en cada una de ellas. Contiene encabezamientos de materias, calificadores, definiciones, referencias cruzadas, sinónimos y listas de términos estrechamente relacionados.
- SNOMED-CT: *Systematized Nomenclature of Medicine – Clinical Terms* es la terminología clínica integral, multilingüe y codificada de mayor amplitud, precisión e importancia desarrollada en el mundo. Es un vocabulario normalizado que permitirá la representación del contenido de los documentos clínicos para su interpretación automática e inequívoca entre sistemas distintos de forma precisa y en diferentes idiomas. Incluye definiciones, términos, relaciones y sinónimos sobre enfermedades, procedimientos clínicos, micro-organismos, síntomas, sustancias y otros conceptos.
- UMLS: *Unified Medical Language System*. Es un conjunto de archivos y software que reúne muchos vocabularios biomédicos y normas para permitir la interoperabilidad entre los sistemas informáticos. Ha sido desarrollado por la *National Library of Medicine* (NLM).

2 Origen y trabajo relacionado

El enfoque en el procesamiento de notas clínicas en la tecnología de información médica podría traer importantes avances en tratamientos médicos y farmacológicos. Por

ello, varias disciplinas como la Informática, la Lingüística y la Biomedicina deben unirse para desarrollar aplicaciones de gestión y búsqueda de recursos médicos.

Muchos sistemas se basan en los textos biomédicos en lengua inglesa. Por ejemplo, *Unified Modeling Language MetaMap Transfer* (UMLS MMTx) (Osborne et al., 2007) es una herramienta configurable comúnmente utilizada por los desarrolladores de sistemas en biomedicina. Creado por investigadores de la *National Library of Medicine* (NLM), MMTx es capaz de identificar los conceptos biomédicos de textos no estructurados y los mapea en los conceptos de UMLS Metathesaurus (Wright et al., 1999). En la Universidad Carlos III, el grupo LABDA ha realizado varias investigaciones centradas en el reconocimiento de entidades médicas, fundamentalmente en la interacción entre medicamentos y centrados principalmente en inglés (Segura-Bedmar, Martínez, and Zazo, 2013; Herrero-Zazo, Segura-Bedmar, and Martínez, 2016) aunque también tienen algunos trabajos en español (Gómez, Bedmar, and Fernández, 2016).

Para el procesamiento de textos biomédicos en español, el Ministerio de Sanidad promueve el uso y aplicación de la ontología SNOMED CT como terminología clínica de referencia para la Historia Clínica Digital Del Sistema Nacional De Salud (HCDSNS). SNOMED-CT se considera la terminología multilingüe más completa del cuidado de la salud en el mundo. En referencia a la ontología SNOMED-CT y el idioma inglés, tenemos varias investigaciones relacionadas (Allones, Martínez, and Taboada, 2014; Sánchez, Batet, and Valls, 2010; Garla and Brandt, 2012).

Como se mencionó anteriormente, las notas clínicas son textos no estructurados, generalmente escritos por especialistas y que presentan características especiales (por ejemplo, a menudo se escriben a mano, contienen errores ortográficos, presentan siglas con múltiples significados y utilizan terminología que viola las convenciones de nomenclatura), por lo que son particularmente difíciles de tratar. Para ello, necesitamos de un procesamiento de textos para la obtención de términos (Barrón-Cedeño et al., 2009; Gelbukh et al., 2010; Bosma and Vossen, 2010).

Vivaldi and Rodríguez (2010) toman como base de información semántica la Wikipedia para crear un sistema de extracción de términos. La metodología consiste en tomar un documento y su correspondiente conjunto de candidatos a términos para comparar los resultados que se obtienen. El sistema se probó en un corpus médico en español y concluyeron que la Wikipedia es un recurso válido para utilizar en esta tarea.

En lo que atañe a trabajos en español aplicando y utlizando la ontología SNOMED-CT, cabe destacar los trabajos de Calleja Ibáñez (2015) y Barahona et al. (2012) donde reconocen de manera automática términos médicos de diferentes documentos. El trabajo de Castro et al. (2010) presenta una anotación semántica de las notas clínicas en español y la aplicación MOSTAS que es una herramienta automatizada para identificar conceptos biomédicos. En la Universidad del País Vasco, Oronoz et al. (2013) desarrollan una herramienta basada en FreeLing para anotar entidades como medicamentos, enfermedades y sustancias de manera automática en textos médicos. Dicha herramienta se conoce como FreeLingMed y aún está en desarrollo. Podemos mencionar también el proyecto de López Rodríguez, Benítez, and Sánchez (2006) Oncoterm, enfocado a un sistema bilingüe de información y recursos oncológicos.

3 Descripción de la investigación propuesta

El objetivo principal de este trabajo consiste en desarrollar herramientas y recursos para el análisis de informes médicos y la extracción de información en documentos clínicos. En principio, nos guiaremos por la descripción detallada que presentan Krauthammer and Nenadic (2004), los cuales elaboran una guía con los pasos a seguir en la extracción de términos en el terreno de la biomedicina:

1. Reconocimiento de términos: denota un conjunto de procedimientos que se usan para reconocer sistemáticamente términos pertinentes en la literatura, es decir, resaltar unidades léxicas que están relacionadas con conceptos de dominio relevantes. Existen varios enfoques para el reconocimiento de términos automático (ATR - *Automatic Term Recognition*):

    - Enfoques basados en diccionario: los métodos basados en diccionario

para el ATR utilizan recursos terminológicos existentes para localizar ocurrencias de término dentro del texto.

- Enfoques basados en reglas: consiste en (normalmente de forma manual) desarrollar reglas que describen estructuras de nomenclatura comunes para ciertas clases de término, buscando indicios ortográficos o léxicos, o características más complejas como las morfo-sintácticas.
- Aprendizaje automático: estos sistemas son aquellos que aprenden y clasifican de forma automática, mejorando con el paso del tiempo. El principal inconveniente de este sistema es preparar un conjunto de entrenamiento inicial para el aprendizaje.
- Enfoques híbridos: combinan diferentes métodos (típicamente los basados en reglas y el aprendizaje automático) y varios recursos (listas de términos específicos, palabras, etc.) para la tarea de reconocimiento de términos.
- Reconocimiento de acrónimos: los términos biomédicos a menudo aparecen de forma acortada o abreviada. Por lo tanto, la capacidad de entender las siglas es obviamente crítica para un sistema de PLN. Encontramos repositorios de acrónimos existentes en el campo biomédico (Chang, Schütze, and Altman, 2002; Rimer and O'Connell, 1998).

2. Clasificación de términos: la tarea de clasificación consiste en desambiguar entre los posibles sentidos de los términos (si hay más de uno) lo que se conoce como desambiguación del sentido del término.
3. Técnicas de mapeo: el objetivo es asignar una ocurrencia de término a una entrada en una fuente de datos de referencia, anotando el término con un ID.

    - Manejo de variabilidad de términos: implica que existan diferentes formas de presentar el mismo concepto, la utilización de abreviaciones, sufijos y prefijos hacen difícil esta tarea.
    - Manejo de ambigüedad de términos: es otro problema importante en la tarea de mapeo de términos; está relacionado con la diversidad de sentidos que se le puede dar a un término con respecto a la fuente de datos utilizada.

4 Metodología y experimentos propuestos

La metodología que se propone para la consecución de este trabajo se presenta a continuación:

1. Estudio y revisión del estado del arte. Se comenzará con el estudio y análisis de la bibliografía existente sobre las técnicas de reconocimiento de entidades nombradas y de las técnicas de mapeo. Se estudiarán las diversas ontologías existentes en el dominio médico tanto en español como en otros idiomas.
2. Adaptación de recursos existentes para poder realizar un análisis de los métodos propuestos.
3. Desarrollo de los recursos y herramientas propios para el análisis y la extracción de información en informes médicos.
4. Implementación de los sistemas que permitan el reconocimiento de entidades médicas en informes médicos en español.
5. Experimentación y evaluación. Se utilizarán los recursos generados para llevar a cabo la experimentación y posteriormente se procederá a la evaluación de los sistemas desarrollados, llevando a cabo una comparación de los resultados obtenidos con los ya existentes. Los resultados obtenidos se pondrán a disposición de la comunidad científica.

5 Conclusión

El objetivo principal del trabajo será la automatización de reconocimiento de entidades en el contenido de informes médicos. Para ello, se aplicarán diversas técnicas de PLN a documentos clínicos escritos en castellano. Esto supone un reto dado que la mayoría de los trabajos realizados en este campo se ha realizado para lengua inglesa y los recursos para el español son bastante limitados.

Cuando se habla de técnicas de PLN nos estamos refiriendo a corrección ortográfica, sistemas de detección de acrónimos, desambiguación, tratamiento de la negación, identificación de conceptos, entre otros. Todas estas tareas se integrarán en sistemas que permitan procesar automáticamente los conceptos encontrados y anotarlos semánticamente con el estándar SNOMED-CT en lengua española.

Agradecimientos

Este trabajo está parcialmente subvencionado por el proyecto REDES (TIN2015-65136-C2-1-R) del MICINN del Gobierno de España.

Bibliografía

Allones, J. L., D. Martínez, y M. Taboada. 2014. Automated mapping of clinical terms into snomed-ct. an application to codify procedures in pathology. *J. Medical Systems*, 38(10):134.

Ananiadou, S. y J. McNaught. 2006. *Text mining for biology and biomedicine*. Artech House London.

Barahona, E. B., I. S. Ramos, A. U. Herradón, A. D. Esteban, y L. P. Morales. 2012. *Procesador automático de informes médicos*.

Barrón-Cedeño, A., G. Sierra, P. Drouin, y S. Ananiadou. 2009. An improved automatic term recognition method for spanish. In *CICLing*, volume 9, pages 125–136. Springer.

Bosma, W. y P. Vossen. 2010. Bootstrapping language neutral term extraction. In *LREC*.

Calleja Ibáñez, P. 2015. *Reconocimiento de enfermedades en fichas técnicas de medicamentos y su anotación con SNOMED-CT*. Ph.D. thesis, ETSI_Informatica.

Castro, E., A. Iglesias, P. Martínez, y L. Castano. 2010. Automatic identification of biomedical concepts in spanish-language unstructured clinical texts. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 751–757. ACM.

Chang, J. T., H. Schütze, y R. B. Altman. 2002. Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association*, 9(6):612–620.

Cohen, A. M. y W. R. Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71.

Garla, V. N. y C. Brandt. 2012. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC bioinformatics*, 13(1):261.

Gelbukh, A., G. Sidorov, E. Lavin-Villa, y L. Chanona-Hernandez. 2010. Automatic term extraction using log-likelihood based comparison with general reference corpus. *Natural Language Processing and Information Systems*, pages 248–255.

Gómez, L. N., I. S. Bedmar, y P. M. Fernández. 2016. Easylecto: Un sistema de simplificación léxica de efectos adversos presentes en prospectos de fármacos en espanol. *Procesamiento del Lenguaje Natural*, 57:177–180.

Hahn, U., M. Romacker, y S. Schulz. 1999. How knowledge drives understanding—matching medical ontologies with the needs of medical language processing. *Artificial Intelligence in Medicine*, 15(1):25–51.

Herrero-Zazo, M., I. Segura-Bedmar, y P. Martínez. 2016. Conceptual models of drug-drug interactions: a summary of recent efforts. *Knowledge-Based Systems*, 114:99–107.

Krauthammer, M. y G. Nenadic. 2004. Term identification in the biomedical literature. *Journal of biomedical informatics*, 37(6):512–526.

López Rodríguez, C. I., P. Benítez, y M. Sánchez. 2006. Terminología basada en el conocimiento para la traducción y la divulgación médicas: el caso de oncoterm. *Panace*, 7(24):228–240.

Oronoz, M., A. Casillas, K. Gojenola, y A. Perez. 2013. Automatic annotation of medical records in spanish with disease, drug and substance names. In *CIARP (2)*, pages 536–543.

Osborne, J. D., S. Lin, L. J. Zhu, y W. A. Kibbe. 2007. Mining biomedical data using metamap transfer (mmtx) and the

unified medical language system (umls). *Gene Function Analysis*, pages 153–169.

Rimer, M. y M. O'Connell. 1998. Bioabacus: a database of abbreviations and acronyms in biotechnology and computer science. *Bioinformatics (Oxford, England)*, 14(10):888–889.

Sánchez, D., M. Batet, y A. Valls. 2010. Web-based semantic similarity: an evaluation in the biomedical domain. *International journal of software and informatics*, 4(1):39–52.

Segura-Bedmar, I., P. Martínez, y M. H. Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Vivaldi, J. y H. Rodríguez. 2010. Using wikipedia for term extraction in the biomedical domain: first experiences. *Procesamiento del Lenguaje Natural*, 45:251–254.

Wright, L. W., H. K. G. Nardini, A. R. Aronson, y T. C. Rindflesch. 1999. Hierarchical concept indexing of full-text documents in the unified medical language system information sources map. *Journal of the Association for Information Science and Technology*, 50(6):514.