Reconocimiento de Entidades Nombradas en español aplicado al dominio biomédico

Named Entity Recognition in Spanish applied to the biomedical domain

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

Resumen: La biomedicina está comenzando a sufrir una gran transformación gracias a la gran cantidad de información almacenada digitalmente. El Procesamiento del Lenguaje Natural tiene el potencial para poder tratar, analizar y comprender el texto. Nuestro principal objetivo consiste en utilizar el reconocimiento de entidades médicas para aplicarlo en diferentes áreas del PLN, de esta manera podremos llegar a conseguir sistemas mucho más potentes y exactos para realizar predicciones. Actualmente, los recursos y herramientas en español existentes no son suficientes, por lo que se estudiaran nuevos métodos para la consecución de este importante reto.
Palabras clave: Reconocimiento de entidades nombradas, terminología médica, ontologías médicas, sistemas de recuperación de información, clasificación de texto, Procesamiento del Lenguaje Natural

Abstract: Biomedicine is undergoing huge transformation thanks to the large amount of information stored digitally. Natural Language Processing has the potential to process, analyze and understand text. Our main objective is to use the recognition of medical entities to apply it in different areas of the PLN, in this way we will be able to achieve more powerful and accurate systems to make predictions. Currently, the existing resources and tools in Spanish are not sufficient, so we will study new methods to achieve this important challenge.
Keywords: Named Entities Recognition, medical terminology, medical ontologies, information retrieval systems, text classification, Natural Language Processing

1 Justificación de la investigación

En el dominio biomédico, la inteligencia artificial permite generar conocimiento y mejorar los servicios sanitarios a partir del tratamiento de información no estructurada que supone una parte importante de los datos que se recogen día a día en la práctica clínica. La inteligencia artificial nos permite aprovechar mejor la información de salud y dar respuesta a los nuevos retos de registro, estructuración y exploración de la información.

Se han hecho esfuerzos considerables para aplicar las tecnologías de minería de texto a la literatura biomédica y los registros clínicos escritos en inglés, pero es cierto que tenemos carencias en cuanto a otros idiomas como el caso del español. Por este motivo, el principal inconveniente que nos encontramos al querer abordar esta tarea es la carencia de corpus disponibles en idiomas distintos al inglés.

Esta tesis está enmarcada dentro del área del Procesamiento del Lenguaje Natural (PLN) en lengua española, concretamente, aborda el estudio de una tarea muy importante dentro de los Sistemas de Recuperación de Información (SRI), como es el Reconocimiento de Entidades Nombradas (NER) (Doan et al., 2012).

El principal objetivo de NER consiste en localizar y clasificar en categorías predefini-

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Sepúlveda-Torres, R. (eds.) *Proceedings of the Doctoral Symposium of the XXXV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*, p. 20–25
Bilbao, Spain, September 25th 2019.

das las entidades encontradas en el texto. Estas categorías pueden variar dependiendo del dominio en el que se aplique. En el caso del dominio biomédico podemos identificar secciones en texto referidas a enfermedades, medicamentos, síntomas, patologías o incluso datos relacionados con pacientes como el nombre, fecha de nacimiento y localización.

El presente trabajo se centra en la identificación de conceptos médicos aplicados en distintas áreas del PLN como son la clasificación de documentos y la recuperación de información (IR por sus siglas en inglés). La clasificación automática de texto consiste en un conjunto de algoritmos, técnicas y sistemas capaces de asignar un documento a una o varias categorías o grupos de documentos (Cohen y Hersh, 2005), por otro lado, la recuperación de información trata de encontrar documentos de una naturaleza no estructurada que satisfacen una necesidad de información (Manning, Raghavan, y Schütze, 2010). Aplicar NER en estas áreas se puede considerar una buena práctica para enriquecer algoritmos y que así estos aprendan de manera más precisa y adecuada para la consecución de los problemas definidos.

En cuanto a los problemas encontrados con el idioma, en los últimos años se han ido creando diferentes competiciones en congresos nacionales que facilitan la obtención de datos para realizar estudios. El Plan de Impulso de las Tecnologías del Lenguaje (Plan TL)[1] tiene como objetivo fomentar el desarrollo del PLN en lengua española y lenguas cooficiales, por otro lado, la Secretaría de Estado para el Avance Digital (SEAD) junto con el Consorcio Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS) trabajan concretamente para el área de biomedicina.

2 Trabajos relacionado

Para dar una visión general, en esta sección se proporcionará una breve descripción de los trabajos relacionados que existen hasta el momento.

En cuanto al reconocimiento de entidades nombradas se han diseñado distintas herramientas automáticas especializadas en el campo de la medicina haciendo uso del PLN, algunas de las más importantes se muestran a continuación:

- LSP-MLP es la primera herramienta desarrollada por la Universidad de Nueva York enfocada en extraer síntomas, medicamentos y efectos secundarios de documentos clínicos.
- MedLEE (*Medical Language Extraction and Encoding System*) (Friedman et al., 1995) utiliza PLN en informes clínicos para ayudar a la toma de decisiones.
- MetaMap (Aronson, 2001) fue desarrollada por la NLM (National Library of Medicine) y mapea en UMLS (*Unified Medical Language System*)[2] los términos médicos encontrados en el texto.
- cTakes (Savova et al., 2010) utiliza un enfoque de aprendizaje automático con métodos basados en reglas.
- MedEx (Xu et al., 2010) se centró principalmente en la identificación de entidades de medicamento.

Los sistemas con el uso de características aprovechan el conocimiento del dominio biomédico para mejorar la clasificación de textos clínicos, Garla (Garla y Brandt, 2012) propone el uso de UMLS para enriquecer el rendimiento de los clasificadores basados en el aprendizaje automático. Por otro lado, Zuccon (Zuccon et al., 2013) también aprovecha la terminología Snomed-CT[3] utilizando conceptos relacionadas con anormalidades y trastornos morfológicos para clasificar informes radiológicos.

En cuanto a los sistemas de recuperación de información (SRI) los enfoques interactivos necesitan la intervención del usuario para refinar y afinar el comportamiento del sistema a fin de obtener una mejor respuesta, mejorando así la relevancia de los elementos recuperados. El sistema propuesto por Mourao (Mourão, Martins, y Magalhães, 2015) permite al usuario añadir nuevas palabras clave tomadas del tesauro MeSH antes de realizar la expansión de la consulta, aunque en este estudio las palabras clave se proponen automáticamente. Esta idea de permitir al usuario refinar la consulta no es nueva, ya que expansión interactiva de la consulta fue propuesta hace más de treinta años (Harman, 1988).

[1] www.plantl.gob.es/

[2] https://www.nlm.nih.gov/research/umls/

[3] http://www.snomed.org/

3 Descripción de la investigación propuesta

La presente tesis se encuentra en proceso de adaptación y generación de recursos para la consecución de las tareas propuestas. El presente año ha estado marcado por diferentes hitos y tareas propuestas por talleres y competiciones existentes en congresos tanto nacionales como internacionales en este área.

El congreso nacional más importante en el área del PLN es la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), este congreso cuenta con un nuevo workshop llamado MEDDOCAN. Por otro lado, contamos con congresos internacionales que abordan la clasificación y reconocimiento de entidades como son BioNLP, CLEF, ACL o TREC.

Todos estos congresos ayudan a seguir unas pautas para desarrollar sistemas, y posteriormente compararte con investigadores que trabajan en mismo campo de estudio. Desde otra perspectiva, estos talleres también aportan nuevas colecciones de texto relacionados con la biomedicina que servirán de prueba para los sistemas. Es por este motivo, que decidimos participar en ellos y así continuar avanzando.

En cuanto a la extracción de información participamos en el congreso *Text REtrieval Conference* (TREC). El objetivo de esta tarea era obtener documentos relevantes en inglés para diferentes consultas que los organizadores te proporcionaban. Las consultas estaban divididas en varios campos relacionados con cánceres y genes. Nuestro objetivo fue obtener los primeros documentos con la consulta original y posteriormente rerankear esos documentos reconociendo entidades médicas en ellos (López-Ubeda et al.,). Para la identificación de conceptos médicos utilizamos la herramienta automática descrita anteriormente llamada MetaMap.

Hemos abordado varias tareas para la clasificación de documentos relacionados con la medicina. La primera fue en *Conference and Labs of the Evaluation Forum* (CLEF) concretamente la tarea eRisk: *Early Detection of Signs of Anorexia*; esta trataba de detectar de forma temprana signos de anorexia en textos extraídos de la plataforma social llamada Reddit[4], para llevar a cabo esta tarea modificamos la matriz de pesos Tf-Idf aportando información obtenida del reconocedor de entidades médicas MetaMap y de la representación de palabras usando word embeddings. De este workshop nació la idea de generar un nuevo recurso en español que estuviera a la disposición de la comunidad científica para realizar experimentos y pruebas acerca de esta enfermedad. Se generó un nuevo corpus extraído de la red social Twitter para detectar problemas de anorexia en textos. Este corpus ha sido generado y presentado para el congreso internacional RANLP 2019 y está compuesto por 5707 tweets clasificados como anorexia y no-anorexia.

Otro workshop aplicando clasificación de texto fue el *Social Media Mining for Health Applications* (SMM4H) en la tarea 1: *Automatic classifications of adverse effects mentions in tweets*, el objetivo fue una clasificación binaria sobre textos en inglés. El sistema diseñado para ello era capaz de distinguir entre los tweets que informan de un Efecto Adverso (EA) y los que no. Para ello aplicamos diferentes estrategias de machine learning y deep learning utilizando redes neuronales convolucionales.

Para la tarea de NER hemos participado en varios workshop. El primero de ellos fue también en SMM4H y la tarea 2: *Extraction of Adverse Effect mentions*, esta incluye la identificación de reacciones adversas a los medicamentos (ADR) reportadas en el texto escrito en inglés. Se utilizó machine learning con *Conditional Random Fields* (Okazaki, 2007) (CRF) con diferentes características obtenidas de word embeddings y clustering de palabras (Brown et al., 1992).

En el congreso SEPLN, MEDDOCAN (Medical Document Anonymization) es la primera tarea dedicada específicamente a la anonimización de documentos médicos en español. Dentro de esta tarea también hemos participado en varias sub-tareas, en la primera de ellas el objetivo era identificar entidades *Protected Health Information* (PHI) y asignarle una categoría a cada entidad. La otra sub-tarea consistía en identificar y ser capaz de enmascarar datos sensibles independientemente del tipo de entidad. Para ambas sub-tareas aplicamos el mismo método que consistía en un enfoque híbrido entre CRF y sistemas basados en reglas. Los resultados obtenidos han sido esperanzadores, alcanzando más del 90 % en ambas tareas.

Para finalizar, en el congreso internacio-

[4] https://www.reddit.com

nal BioNLP, hemos participado en un nuevo workshop llamado PharmaCoNER (*Pharmacological Substances, Compounds and proteins and Named Entity Recognition*). Este workshop está dedicado al reconocimiento de químicos y medicamentos en textos médicos en español y está compuesto por varias subtareas. Por un lado, aborda un problema de NER donde debemos encontrar la entidad nombrada junto con una etiqueta asociada; por otro lado, conlleva asignarle un código Snomed-CT a las entidades nombradas identificadas anteriormente. El resultado obtenido fue el esperado, en la primera sub-tarea se desarrollaron enfoques de machine learning y deep learning y le añadimos información adicional con la ayuda de la terminología Snomed-CT, esto contribuyó a los resultados ya que conseguimos detectar y anotar el 78 % de los conceptos. Para la segunda sub-tarea, se generó una arquitectura basada en diccionarios para encontrar el identificador de Snomed-CT más apropiado.

4 Metodología propuesta

La metodología de este trabajo se basa en el estudio del estado del arte de las distintas áreas en las que estamos trabajando como son: la recuperación de información, el reconocimiento de entidades nombradas y la clasificación dentro del dominio biomédico. Este estudio consistirá en una revisión bibliográfica de ya lo existente, centrándonos principalmente en el uso de las entidades nombradas médicas para enriquecer sistemas de clasificación y de recuperación de información.

Para la consecución de la tarea de identificación de conceptos médicos se crearán nuevos sistemas basados en diccionarios, en reglas, en aprendizaje automático o híbridos.

En los sistemas basados en diccionarios es necesario partir de un vocabulario controlado, en el campo de la medicina contamos con ontologías, terminologías, diccionarios y lexicones. Algunos de los vocabularios más populares hasta el momento son UMLS, Snomed-CT, ICD-10 (*International Statistical Classification of Diseases*) o MeSH (*Medical Subject Headings*)[5]. Con estos recursos conseguimos ir un paso más allá, pudiendo mapear el concepto identificado en un código estandarizado.

Los métodos basados en diccionario utilizan recursos terminológicos existentes para localizar ocurrencias de término dentro del texto. Los sistemas basados en reglas desarrollan patrones en el texto que describen estructuras de nomenclatura comunes para ciertos términos, buscando indicios ortográficos o léxicos, o características más complejas como las morfo-sintácticas (Chen et al., 2019; Nguyen et al., 2010). Los 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.

Por otro lado, se estudiarán los diferentes métodos de aprendizaje automático (Zhou et al., 2004) y aprendizaje profundo (GuoDong y Jian, 2004) existentes tanto para el área de NER como para la clasificación de documentos.

Otra tarea a abordar será el uso de técnicas novedosas con Modelos de Lenguaje. Los modelos de lenguaje son el principal problema para algunas tareas del Procesamiento de Lenguaje Natural tales como los sistemas conversacionales, generación de textos o resúmenes de textos. Pero también ha sido aplicado a otros ámbitos como la identificación de términos. Un modelo de lenguaje entrenado aprende la probabilidad de que ocurra una palabra basándose en la secuencia anterior de palabras usadas en el texto. Los modelos de lenguaje pueden ser operados a nivel de caracteres, a nivel de n-gramas, a nivel de frases o incluso a nivel de párrafos. Un ejemplo es el de Rizwan (Parvez et al., 2018), el modelo que generó aprende la distribución de probabilidad sobre todas las palabras candidatas aprovechando la información del tipo de entidad.

Existen numerosos modelos de lenguaje pre-entrenados actualmente como por ejemplo: *Embeddings from Language Models* (ELMo) (Peters et al., 2018), *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2018), *Universal Language Model Fine-tuning* (ULMFiT) (Howard y Ruder, 2018) o BioBERT que es un modelo de representación del lenguaje para el dominio biomédico, especialmente diseñado para tareas de minería de textos biomédicos, como el reconocimiento biomédico de entidades nombradas.

Finalmente, se realizará una experimentación y evaluación. Se probarán los sistemas

[5]https://meshb.nlm.nih.gov/

desarrollados para ver su efectividad, de esta manera, podremos comparar nuestros resultados. Los resultados logrados será compartidos y se pondrán a disposición de la comunidad científica.

5 Elementos de investigación para discusión

La clasificación, recuperación de información, anotación e identificación de entidades es un tema de interés en el PLN, nuestra intención en este trabajo es discutir las siguientes aspectos para seguir profundizando en el estudio:

- ¿Es necesario la creación de recursos para el NER médicas en español?
- ¿Qué sistemas son los principalmente usados en el NER en el dominio biomédico?
- ¿Basta utilizar los vocabularios controlados anteriormente mencionados para mejorar la clasificación o la recuperación de información?
- ¿Qué técnicas se utilizan para enriquecer con los vocabularios médicos existes?
- ¿Qué otros recursos contribuyen a la consecución del NER dentro del campo de la medicina?
- ¿Cuánta información extra aporta el reconocimiento de entidades en el área de la clasificación? ¿Y para la recuperación de información?

Agradecimientos

Este trabajo está parcialmente subvencionado por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto LIVING-LANG (RTI2018-094653-B-C21) y el proyecto REDES (TIN2015-65136-C2-1-R) del Gobierno de España.

Bibliografía

Aronson, A. R. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. En *Proceedings of the AMIA Symposium*, página 17. American Medical Informatics Association.

Brown, P. F., P. V. Desouza, R. L. Mercer, V. J. D. Pietra, y J. C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Chen, L., L. Song, Y. Shao, D. Li, y K. Ding. 2019. Using natural language processing to extract clinically useful information from chinese electronic medical records. *International journal of medical informatics*, 124:6–12.

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

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doan, S., N. Collier, H. Xu, P. H. Duy, y T. M. Phuong. 2012. Recognition of medication information from discharge summaries using ensembles of classifiers. *BMC medical informatics and decision making*, 12(1):36.

Friedman, C., S. B. Johnson, B. Forman, y J. Starren. 1995. Architectural requirements for a multipurpose natural language processor in the clinical environment. En *Proceedings of the Annual Symposium on Computer Application in Medical Care*, página 347. American Medical Informatics Association.

Garla, V. N. y C. Brandt. 2012. Ontology-guided feature engineering for clinical text classification. *Journal of biomedical informatics*, 45(5):992–998.

GuoDong, Z. y S. Jian. 2004. Exploring deep knowledge resources in biomedical name recognition. En *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, páginas 96–99. Association for Computational Linguistics.

Harman, D. 1988. Towards interactive query expansion. páginas 321–331. cited By 110.

Howard, J. y S. Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

López-Ubeda, P., M. C. Dıaz-Galiano, M.-T. Martın-Valdivia, y L. A. Urena-López. Using clustering to filter results of an information retrieval system.

Manning, C., P. Raghavan, y H. Schütze. 2010. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103.

Mourão, A., F. Martins, y J. Magalhães. 2015. Multimodal medical information retrieval with unsupervised rank fusion. *Computerized Medical Imaging and Graphics*, 39:35–45.

Nguyen, A. N., M. J. Lawley, D. P. Hansen, R. V. Bowman, B. E. Clarke, E. E. Duhig, y S. Colquist. 2010. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association*, 17(4):440–445.

Okazaki, N. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Parvez, M. R., S. Chakraborty, B. Ray, y K.-W. Chang. 2018. Building language models for text with named entities. *arXiv preprint arXiv:1805.04836*.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, y L. Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, y C. G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Xu, H., S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, y J. C. Denny. 2010. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

Zhou, G., J. Zhang, J. Su, D. Shen, y C. Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190.

Zuccon, G., A. S. Wagholikar, A. N. Nguyen, L. Butt, K. Chu, S. Martin, y J. Greenslade. 2013. Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology. *AMIA Summits on Translational Science Proceedings*, 2013:300.