Integración de Conocimiento para la Mejora de Sistemas de Recuperación de Información

Knowledge Integration for Improving Information Retrieval Systems

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

Resumen: Con el paso del tiempo, está tomando más importancia el intercambio y manejo de la información, sobre todo el ámbito biomédico, pues estos documentos contienen información relevante sobre síntomas, enfermedades, alergias, etc. Por ello, se necesitan sistemas para poder tratar dicha información de manera adecuada. Este trabajo se enmarca dentro del área del Procesamiento del Lenguaje Natural en lengua española, concretamente, aborda el estudio de tareas tan importantes dentro de los Sistemas de Recuperación de Información, como son el Reconocimiento de Entidades Nombradas o la integración de conocimiento desde fuentes externas. En nuestro caso, propondremos identificar y clasificar elementos en un informe clínico estudiando diccionarios y ontologías en el dominio biomédico y diferentes idiomas, algoritmos y recursos existentes. Finalmente, crearemos nuevos sistemas para posteriormente probarlos y ponerlos a disposición de la comunidad científica.
Palabras clave: Recuperación de información, reconocimiento de entidades, UMLS, SNOMED-CT, ICD10, cTAKES, MetaMap, aprendizaje automático

Abstract: Over time, the exchange and management of information is becoming more important, especially in the biomedical field. These documents contain relevant information on symptoms, diseases, allergies, etc. For this reason, we need systems to be able to process this information properly. This work is framed within the area of Natural Language Processing in Spanish language, specifically, the study of a very important task within the Information Retrieval Systems, such as the Named Entities Recognition and the knowledege integration from external sources. In our case, we will propose the identification and classification of medical concepts in clinical reports by studying dictionaries and ontologies in the biomedical domain and different languages, algorithms and existing resources. Finally, we will create new systems to later test them and make them available to the scientific community.
Keywords: Information retrieval, entity recognition, UMLS, SNOMED-CT, ICD10, cTAKES, MetaMap, machine learning

1 Introducción

Unos de los retos tecnológicos que se plantean en esta tesis es la exploración de la Web Semántica haciendo uso de herramientas y recursos de Procesamiento del Lenguaje Natural (PLN). El objetivo es entender automáticamente el lenguaje humano apoyándonos en la inteligencia artificial. Se centrará concretamente en la búsqueda y recuperación de información para poder obtener respuestas completas, correctas y oportunas a las necesidades de información de los usuarios.

Se hace necesario el uso de herramientas que faciliten el acceso y recuperación de datos, de ahí nacen los llamados sistemas de recuperación de información. Un Sistema de Recuperación de Información (SRI) se puede definir como un proceso capaz de almacenar, recuperar y mantener información (Kowalski, 2007). Estos sistemas amplían el espectro de cobertura en la búsqueda a partir de bases de datos documentales, además de poder incorporar diversos métodos para el ordenamiento de los documentos que mejoren la relevancia de los resultados para el usuario. Dentro de los SRI la gestión de informes médicos está obteniendo gran

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Moreno, I. (eds.) *Proceedings of the Doctoral Symposium of the XXXIV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, p. 31–36 Sevilla, Spain, September 19th 2018.

importancia, pues los datos contenidos en dichos informes son relevantes tanto para los enfermos como para los especialistas en medicina. Los informes clínicos contienen información sobre el paciente, medicación, resultados de análisis, diagnósticos, dosis, etc. Manteniendo esa información digitalizada obtenemos grandes ventajas, se reduce el tiempo de trabajo del personal de salud y se mejora la calidad de la atención entre otros.

El cumplimiento de ciertas normas necesarias para desarrollar de manera coherente la recuperación de contenidos web supone la creación de ontologías sobre el dominio o área de conocimiento específico. "*Una Ontología define los términos básicos y las relaciones entre ellos de un tema en concreto como también la reglas para combinarlos y extender otros términos y relaciones del vocabulario*" (Neches et al., 1991). Gruber (1995) también aportó conocimiento en base a las ontologías para obtener conocimiento a partir de ellas y compartirlo "*se puede considerar una ontología como un sistema de representación del conocimiento en un ámbito específico que puede organizarse en forma jerárquica para facilitar la representación y comprensión del conocimiento*". En este trabajo estudiaremos el diccionario ICD10 (*International Classification of Diseases*) y las ontologías UMLS (*Unified Medical Language System*) y SNOMED-CT (*Systematized Nomenclature of Medicine – Clinical Terms*) para resolver los problemas antes mencionados.

Para llegar a obtener documentos que satisfagan las necesidades del usuario en los SRI, haremos uso de diferentes subtareas como la identificación de términos o la clasificación, donde ellos se convierten en la clave para acceder a documentación bibliográfica y literatura relacionada. El objetivo del Reconocimiento de Entidades Nombradas (NER *Named Entity Recognition* por sus siglas en inglés) es identificar en un texto menciones de elementos pertenecientes a una determinada clase de conceptos.

Aunque el trabajo está centrado en el dominio biomédico, la idea principal es crear técnicas y algoritmos lo suficientemente flexibles para ser aplicados a distintos ámbitos con éxito. La mayor dificultad radicaría en encontrar recursos de calidad para el dominio donde se quieran aplicar las técnicas desarrolladas.

2 Origen y trabajo relacionado

En los últimos años, las ontologías han desempeñado un papel importante en el campo biomédico (Rubin, Shah, y Noy, 2007). Se han utilizado ontologías de dominio biomédico para la anotación de datos, la integración de información, el descubrimiento de conocimientos y otras aplicaciones.

Existe una gran cantidad de ontologías en este ámbito y muchas de ellas estrechamente relacionadas, por lo tanto, las calidades de dichas ontología pueden variar mucho. Encontramos investigaciones interesantes y de diversa índole sobre las ontologías más comúnmente utilizadas como: SNOMED-CT (Stearns et al., 2001; Patrick, Wang, y Budd, 2007), UMLS (Bodenreider, 2004; Huang et al., 2005; Brennan y Aronson, 2003) y CIE-10 (Névéol et al., 2017).

En cuanto a los trabajos relacionados con el reconocimiento de entidades médicas en inglés, encontramos sistemas de detección automática, como son MetaMap y cTAKES, ambos se encuentran disponibles y permiten descubrir entidades médicas en texto e identificarlas como conceptos ontológicos.

MetaMap es creado por investigadores de la *National Library of Medicine* (NLM) y es capaz de identificar los conceptos biomédicos de textos no estructurados y los mapea en los conceptos de UMLS Metathesaurus (Aronson, 2001), por otro lado, Apache cTAKES (*Apache Clinical Text Analysis and Knowledge Extraction System*) (Savova et al., 2010) es un sistema de procesamiento de lenguaje natural para extracción de información en texto clínico. Rodríguez González et al. (2015) realiza una comparativa entre las dos herramientas sobre MedLinePlus [1].

Existen otras alternativas para el inglés, como MedLEE (*Medical Language Extraction and Encoding System*) originalmente desarrollado para radiología y posteriormente extendido a otros subdominios. BioPortal (Noy et al., 2009; Zheng et al., 2018) del NCBO (*National Center for Biomedical Ontologies*) representa mapeos entre términos de diferentes ontologías, actualmente incorpora más de 600 ontologías.

En cuanto al idioma español existen menos herramientas disponibles, si bien podemos

[1] https://medlineplus.gov/

encontrar algunos ejemplos como la versión en español de MetaMap (Carrero, Cortizo, y Gómez, 2008) y Freeling-Med (Oronoz et al., 2013).

3 Investigación propuesta

Este trabajo de tesis se encuentra en fase de desarrollo y adaptación de recursos existentes. Por lo que a lo largo de este año se han ido siguiendo una serie de hitos marcados por los actuales talleres y competiciones que existen dentro del dominio médico como son los talleres TASS e IberEval que se celebran en el marco del congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), el foro CLEF eHealth y la competición TREC; con todos ellos pretendemos desarrollar nuestros sistemas, evaluarlos y crear una comparativa para seguir avanzando.

En cuanto a reconocimiento de entidades médicas en español, se ha diseñado una herramienta propia para la detección de términos biomédicos en un texto llamada BSB[2] (Buscador Semántico Biomédico) (López-Úbeda et al., 2018c). Las bases de conocimiento usadas en el reconocedor son UMLS, SNOMED e ICD10, todas en ellas en español. La herramienta utiliza la biblioteca NLTK (*Natural Language Toolkit*) desarrollada en el lenguaje de programación Python y el analizador sintáctico incluido en la herramienta CoreNLP en español (Manning et al., 2014) para obtener una mayor precisión a la hora de identificar terminología (lematización, desambiguación, palabras compuestas).

Pretendemos modelar el lenguaje humano en escenarios de dominio biomédico para que los documentos electrónicos puedan ser legibles por máquinas desde un punto de vista semántico, para ello, participamos en la Task 3. eHealth Knowledge Discovery del Taller de Análisis Semántico de la SEPLN (TASS) (López-Úbeda et al., 2018b). Con esta tarea pudimos crear un sistema donde identificar todas las frases clave en un documento escrito en español y asignar una etiqueta (concepto o acción) a todas aquellas frases clave detectadas. Para ello, adaptamos la herramienta desarrollada BSB.

Por otra parte, también hemos intentado estudiar otros idiomas como el francés. En la Task 1, Multilingual Information Extraction - ICD10 coding (Névéol et al., 2018) del CLEF eHealth, el principal objetivo es crear un sistema basado en técnicas de PLN para la detección de códigos ICD10 utilizando diferentes algoritmos de aprendizaje automático. Primero, encontramos todos los posibles códigos ICD10 mencionados en el texto y a continuación, creamos varias medidas para tratar el texto del concepto identificado. Con estas métricas entrenamos diferentes algoritmos de aprendizaje automático y elegimos el mejor modelo a utilizar en nuestro sistema.

Un reto marcado por el taller DIANN (*Disability annotation on documents from the biomedical domain*) incluído en la SEPLN ha sido el anotar discapacidades encontradas en un texto. Se trata como reto porque las herramientas para la detección de entidades nombradas en el ámbito biomédico no consideran las discapacidades como un concepto distintivo, sino como cualquier otro signo. Por lo tanto, no permiten distinguir una discapacidad, generalmente una condición permanente, de otros signos asociados a enfermedades. Para esta tarea utilizamos nuestro sistema BSB para los textos en español y MetaMap con UMLS para el inglés y pudimos ver las diferencias existentes. Además, para el español, incorporamos una nueva fuente de conocimiento basada en siglas que incluía terminología sobre discapacidades (López-Úbeda et al., 2018a).

En cuanto a trabajos relacionados con la obtención de documentos relevantes en SRI hemos participado en la Task3 - Consumer Health Search Task (Jimmy et al., 2018), donde aplicamos la técnica de expansión de consultas utilizando el buscador Google. Identificamos los conceptos médicos en los resultados de Google utilizando cTAKES para evitar introducir ruido en la consulta con conceptos que no sean de dominio médico. Por otro lado, en fase de desarrollo se encuentra la participación en TREC Precision Medicine Track para 2018, basada en su anterior tarea (Roberts et al., 2017) donde estamos aplicando técnicas de *word embedding*(Mikolov et al., 2013) para expandir la consulta original con terminología cercana a los conceptos más importantes de la consulta.

[2]http://sinai.ujaen.es/demo/bsb/

4 Metodología propuesta

Para el desarrollo de esta tesis se propone la siguiente metodología a seguir:

1. Estudio y revisión del estado del arte. Se comenzará con el estudio y análisis de la bibliografía existente sobre los sistemas de recuperación de información utilizando la técnica de reconocimiento de entidades.
2. Estudiar las diversas ontologías existentes tanto en español como en otros idiomas.
3. Adaptar los recursos existentes para poder realizar un análisis de los métodos propuestos.
4. Desarrollo de recursos y herramientas propios para el análisis y la extracción de información en informes médicos.
5. Implementación de los sistemas que permitan satisfacer las necesidades de información de un usuario.
6. 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 Elementos específicos para discusión

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

1. ¿Qué ontologías médicas son las más utilizadas y por qué?
2. ¿Cuáles son los algoritmos y recursos para hacer búsquedas más aproximadas en los SRI?
3. ¿Qué tipo de información de una ontología puede mejorar un SRI? ¿Qué aporta esa información (precisión, cobertura, diversidad, etc.)?
4. ¿Qué sistemas automáticos existen actualmente para identificar entidades médicas en un texto? ¿Y específico para el idioma español?
5. ¿Es necesario crear un reconocedor de entidades médicas multilingüe?
6. Comparativa entre los NER en los idiomas inglés y español.

Agradecimientos

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

Bibliografía

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

Bodenreider, O. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Brennan, P. F y Alan R Aronson. 2003. Towards linking patients and clinical information: detecting umls concepts in e-mail. *Journal of Biomedical Informatics*, 36(4-5):334–341.

Carrero, F, José Carlos Cortizo, y José María Gómez. 2008. Building a spanish mmtx by using automatic translation and biomedical ontologies. En *International Conference on Intelligent Data Engineering and Automated Learning*, páginas 346–353. Springer.

Gruber, T. R. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-computer Studies*, 43(5-6):907–928.

Huang, Y, Henry J Lowe, Dan Klein, y Russell J Cucina. 2005. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the umls specialist lexicon. *Journal of the American Medical Informatics Association*, 12(3):275–285.

Jimmy, Guido Zuccon, Joao Palotti, Lorraine Goeuriot, y Liadh Kelly. 2018. Overview of the clef 2018 consumer health search task. En *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*.

Kowalski, G. J. 2007. *Information retrieval systems: theory and implementation*, volumen 1. Springer.

López-Úbeda, P, Manuel Carlos Díaz Galiano, María Teresa Martín-Valdivia, y Salud Jiménez-Zafra. 2018a. Sinai at diann-ibereval 2018. annotating disabilities in multi-language systems with umls. En *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*.

López-Úbeda, P, Manuel Carlos Díaz-Galiano, María Teresa Martín-Valdivia, y Luis Alfonso Ureña López. 2018b. SINAI en TASS 2018 task 3. clasificando acciones y conceptos con UMLS en medline (SINAI in TASS 2018 task 3. classifying actions and concepts with UMLS on medline). En *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, páginas 77–82.

López-Úbeda, P, Manuel Carlos Díaz Galiano, Arturo Montejo Ráez, Fernando Martínez Santiago, Alberto Andreu-Marín, Martín, María Teresa, y Luis Alfonso Ureña López. 2018c. Buscador semántico biomédico. *Procesamiento del Lenguaje Natural*, 61:189–192.

Manning, C, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, y David McClosky. 2014. The stanford corenlp natural language processing toolkit. En *Proceedings of 52nd annual meeting of the Association for Computational Linguistics, ACL: system demonstrations*, páginas 55–60.

Mikolov, T, Kai Chen, Greg Corrado, y Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Neches, R, Richard E Fikes, Tim Finin, Thomas Gruber, Ramesh Patil, Ted Senator, y William R Swartout. 1991. Enabling technology for knowledge sharing. *AI magazine*, 12(3):36.

Névéol, A, A Robert, F Grippo, C Morgand, C Orsi, L Pelikán, L Ramadier, G Rey, y P Zweigenbaum. 2018. Clef ehealth 2018 multilingual information extraction task overview: ICD10 coding of death certificates in french, hungarian and italian. En *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*.

Névéol, A, Robert N Anderson, K Bretonnel Cohen, Cyril Grouin, Thomas Lavergne, Grégoire Rey, Aude Robert, Claire Rondet, y Pierre Zweigenbaum. 2017. Clef ehealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in english and french. En *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*, página 17.

Noy, N. F, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, y Christopher G Chute. 2009. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(suppl_2):W170–W173.

Oronoz, M, Arantza Casillas, Koldo Gojenola, y Alicia Perez. 2013. Automatic annotation of medical records in spanish with disease, drug and substance names. En *Iberoamerican Congress on Pattern Recognition*, páginas 536–543. Springer.

Patrick, J, Yefeng Wang, y Peter Budd. 2007. An automated system for conversion of clinical notes into snomed clinical terminology. En *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68*, páginas 219–226. Australian Computer Society, Inc.

Roberts, K, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, y Shubham Pant. 2017. Overview of the trec 2017 precision medicine track. En *Text REtrieval Conference, TREC, Gaithersburg, MD*.

Rodríguez González, A, Roberto Costumero Moreno, Marcos

Martínez Romero, Mark Denis Wilkinson, y Ernestina Menasalvas Ruiz. 2015. Extracting diagnostic knowledge from medline plus: a comparison between metamap and ctakes approaches. *Current Bioinformatics*, 375:1–7.

Rubin, D. L, Nigam H Shah, y Natalya F Noy. 2007. Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, 9(1):75–90.

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

Stearns, M. Q, Colin Price, Kent A Spackman, y Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. En *Proceedings of the AMIA Symposium*, página 662. American Medical Informatics Association.

Zheng, L, Yan Chen, Gai Elhanan, Yehoshua Perl, James Geller, y Christopher Ochs. 2018. Complex overlapping concepts: An effective auditing methodology for families of similarly structured bioportal ontologies. *Journal of Biomedical Informatics*, 83:135 – 149.