Clasificación e identificación de conceptos biomédicos en texto libre

Naiara Perez[1*], Montse Cuadros[1] y German Rigau[2]

[*]Autor de correspondencia: nperez@vicomtech.org
[1]Grupo HSLT, Vicomtech, Paseo Mikeletegi, 57, Donostia/San Sebastián
Lista completa de autores disponible al final del artículo

Una de las mayores limitaciones del big-data en el dominio de la salud es que al menos 80 % de los documentos no son aptos para su procesamiento automático sin previo tratamiento por consistir en texto libre [1]. En este trabajo presentamos la primera herramienta que permite estructurar texto libre clínico mediante terminologías de referencia y técnicas de Procesamiento del Lenguaje Natural en español evaluada y comparada con herramientas de referencia.

En concreto, la herramienta que proponemos es capaz de clasificar e identificar conceptos médicos en texto clínico escrito. El desarrollo actual permite la anotación de conceptos médicos como un problema de búsqueda de correspondencias léxicas en el compendio de terminologías UMLS [2] y de desambiguación de acepciones. La herramienta cuenta, entre otros, con módulos de detección y resolución de abreviaturas biomédicas y de análisis morfosintáctico. Además, es altamente configurable ya que permite al usuario elegir qué terminologías se deben utilizar y qué tipos de conceptos (p.e. síntomas y enfermedades, medicamentos, conceptos temporales) se deben encontrar en los textos.

Esta herramienta se ha evaluado con un corpus de referencia llamado Mantra GSC [3], que consta de 200 documentos biomédicos anotados a nivel conceptual. La puntuación F-1 de la clasificación y de la identificación de conceptos es de 0,65 y 0,62 respectivamente. Estos resultados son los primeros reportados para una herramienta en el análisis de textos de informes médicos para el español, y sugieren que la herramienta es comparable a las más referidas en la literatura, las cuales sólo procesan textos en inglés (p.e., [4, 5, 6]).

Las anotaciones que se obtienen gracias a las terminologías actualmente existentes y creadas por expertos componen la base sobre la que desarrollar aplicaciones más complejas (p.e. de codificación automática) o complementar las ya existentes (p.e. de Soporte a la Decisión).

Detalles sobre los autores
[1]Grupo HSLT, Vicomtech, Paseo Mikeletegi, 57, Donostia/San Sebastián. [2]IXA Taldea, EHU/UPV, Paseo Manuel Lardizabal, 1, Donostia/San Sebastián.

Referencias

1. Färber, M.: Semantic Search for Novel Information vol. 31, (2017)
2. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), 267–270 (2004)
3. Kors, J.A., Clematide, S., Akhondi, S.A., van Mulligen, E.M., Rebholz-Schuhmann, D.: A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. Journal of the American Medical Informatics Association **22**(5), 948–956 (2015)
4. Aronson, A.R.: MetaMap: Mapping Text to the UMLS Metathesaurus. Bethesda, MD: NLM, NIH, DHHS (2006)
5. Dai, M., Shah, N.H., Xuan, W., Musen, M.A., Watson, S.J., Athey, B.D., Meng, F., et al.: An Efficient Solution for Mapping Free Text to Ontology Terms. AMIA Summit on Translational Bioinformatics **21** (2008)
6. Jonquet, C., Shah, N.H., Musen, M.A.: The Open Biomedical Annotator. AMIA Summit on Translational Bioinformatics **2009** (2009)