Título: DESARROLLO DE HERRAMIENTAS INTELIGENTES PARA VISTAS ESPECIALIZADAS E INTEGRADAS DE LA BIBLIOGRAFÍA CLÍNICA

Nombre: Pérez Pérez, Martín

Universidad: Universidad de Vigo

Departamento: Informática

Fecha de lectura: 11/03/2019

Programa de doctorado: Programa de Doctorado en Sistemas Software Inteligentes y Adaptables por la Universidad de Vigo

Dirección:

> **Director:** Martin Krallinger

> **Director:** Anália Maria Garcia Lourenço

Tribunal:

> **presidente:** FERNANDO DÍAZ GÓMEZ

> **secretario:** MARÍA REYES PAVÓN RIAL

> **vocal:** GILBERTO IGREJAS

Descriptores:

> SOFTWARE

> INFORMATICA

> BASES DE DATOS

> ENFERMEDADES INFECCIOSAS

El fichero de tesis no ha sido incorporado al sistema.

Resumen: En los últimos años, los avances en las ciencias de la vida han causado un aumento considerable del número de estudios biomédicos publicados. Varios portales han sido creados para facilitar el acceso a las publicaciones científicas, sin embargo debido al ritmo acelerado con el que se publican nuevos artículos, es cada vez más difícil recopilar información en tiempo útil [1]. El aumento de publicaciones se debe, en parte, al cambio tecnológico producido en el proceso de producción científica por medio de las publicaciones electrónicas y revistas online, junto con un grado, cada vez mayor, de especialización científica (y por tanto de las publicaciones especializadas).La minería de textos está demostrando tener el potencial para ayudar en la recopilación y el análisis de documentos científicos, dado que proporciona un enfoque válido para el tratamiento automatizado y eficiente de grandes colecciones bibliográficas [2]–[4].
El proceso de obtención de conocimiento a través de modelos de procesamiento del lenguaje natural y clasificación automática de contenidos depende, en gran medida, de la capacidad de producir corpora anotados de gran calidad. Un corpora anotado tiene un gran valor en el ámbito del análisis lingüístico, sin embargo, su preparación es costosa en cuanto a tiempo y recursos se refiere. En este este sentido, existen cada vez más iniciativas de la comunidad, a nivel internacional, que fomentan la creación de corpora anotados para la

evaluación controlada de los algoritmos de procesamiento del lenguaje natural. Estas iniciativas están centradas en distintos ámbitos específicos, pero tienen una especial relevancia en el campo de la biomedicina. Iniciativas internacionales como BioCreative [5], BioNLP [6], o CALBC [7] ponen a disposición de los desarrolladores recursos y herramientas para evaluar el rendimiento de los modelos de análisis lingüístico, creados en base a distintas directrices elaboradas por expertos en el dominio [8], [9]. Estas iniciativas atraen una participación significativa de la comunidad internacional entre los que destacan: expertos en el procesamiento del lenguaje natural, el aprendizaje automatizado (también llamado aprendizaje de máquina) y la minería de textos. Los sistemas de evaluación y los corpora presentados son de vital importancia a la hora de desarrollar y mejorar el rendimiento de los sistemas actuales [10].

Otra parte crucial para la mejora del acceso eficiente a la información científica a gran escala es la representación y el resumen adecuado de los contenidos extraídos de manera automática. Así pues, aunque los modelos automáticos presenten un buen rendimiento, si la información no se organiza y se completa de acorde a la temática estudiada y al volumen manejado, es muy probable que el usuario no tenga acceso a los contenidos de manera eficiente. Está demostrado que la obtención de datos estructurados centrados en un campo específico de la literatura clínica, puede servir para validar estudios científicos e incluso puede ayudar a esbozar nuevas hipótesis [11], [12]. Las ontologías, los vocabularios controlados y las bases de datos curadas1 manualmente son algunas de las herramientas que se están desarrollando para ayudar a los científicos y profesionales a organizar y acceder fácilmente a la gran cantidad de información que permanece oculta en la literatura. En este sentido, la reconstrucción de redes de información se postula como un soporte intuitivo, dinámico y escalable capaz de dar soporte a la minería de eventos o asociaciones semánticas relevantes, entre los contenidos, permitiendo así nuevas metodologías de análisis y clasificación de documentos.

Este trabajo de investigación explora distintas metodologías para: (i) mejorar sistemáticamente la calidad final de los corpora, (ii) establecer una evaluación continuada del rendimiento de los modelos de reconocimiento automático (también denominados NERs por las siglas en inglés Named Entity Recognisers) y (iii) representar de manera eficiente los contenidos curados de forma manual o de forma semiautomática [13].

[1]W. W. M. Fleuren and W. Alkema, “Application of text mining in the biomedical domain.,” Methods, vol. 74, pp. 97–106, Mar. 2015.

[2]Z. Javed and H. Afzal, “Biomedical text mining for concept identification from traditional medicine literature,” in 2014 International Conference on Open Source Systems & Technologies, 2014, pp. 206–211.

[3]D. Piedra, A. Ferrer, and J. Gea, “Text mining and medicine: usefulness in respiratory diseases.,” Arch. Bronconeumol., vol. 50, no. 3, pp. 113–9, Mar. 2014.

[4]S. Zaremba et al., “Text-mining of PubMed abstracts by natural language processing to create a public knowledge base on molecular mechanisms of bacterial enteropathogens.,” BMC Bioinformatics, vol. 10, p. 177, Jan. 2009.

[5]C. N. Arighi et al., “Overview of the BioCreative III Workshop.,” BMC Bioinformatics, vol. 12 Suppl 8, no. 8, p. S1, Jan. 2011.

[6]The Association for Computational Linguistics, “Proceedings of the 4th BioNLP Shared Task Workshop,” 2016.

[7]D. Rebholz-Schuhmann et al., “Assessment of NER solutions against the first and second CALBC Silver Standard Corpus,” J. Biomed. Semantics, vol. 2, no. Suppl 5, p. S11, Oct. 2011.

[8]M. Krallinger et al., “The CHEMDNER corpus of chemicals and drugs and its annotation principles.,” J.

Cheminform., vol. 7, no. Suppl 1 Text mining for chemistry and the CHEMDNER track, p. S2, 2015.
[9]F. Leitner, M. Krallinger, G. Cesareni, and A. Valencia, “The FEBS Letters SDA corpus: a collection of protein interaction articles with high quality annotations for the BioCreative II.5 online challenge and the text mining community.,” FEBS Lett., vol. 584, no. 19, pp. 4129–30, Oct. 2010.
[10]K. M. Hettne, A. J. Williams, E. M. van Mulligen, J. Kleinjans, V. Tkachenko, and J. A. Kors, “Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining.,” J. Cheminform., vol. 2, no. 1, p. 3, Jan. 2010.
[11]R. Rodriguez-Esteban, “Biomedical text mining and its applications.,” PLoS Comput. Biol., vol. 5, no. 12, p. e1000597, 2009.
[12]F. Zhu et al., “Biomedical text mining and its applications in cancer research,” J. Biomed. Inform., vol. 46, no. 2, pp. 200–211, Apr. 2013.
[13]D. Howe et al., “The future of biocuration,” Nature, vol. 455, no. 7209, pp. 47–50, Sep. 2008.