Tecnologías de lenguaje natural en el sector sanitario y en la Administración.

El Procesamiento del Lenguaje Natural (PLN) es un campo de la Inteligencia Artificial (IA) relacionado con la lingüística computacional que se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para el análisis y generación del lenguaje natural, entendiendo como tal las lenguas que hablamos los humanos.

RAQUEL MARTÍNEZ
Profesora Catedrática de Universidad de la Universidad Nacional de Educación a Distancia

Mª SOTO MONTALVO
Profesora Contratada Doctora de la Universidad Rey Juan Carlos

1. CONTEXTO DE LAS TECNOLOGÍAS DEL LENGUAJE EN LA ADMINISTRACIÓN

La denominación *tecnologías de lenguaje humano* es equivalente. Desde los años 50 se investiga para que los ordenadores sean capaces de procesar el lenguaje humano, comprender su significado y generar lenguaje.

Este campo de investigación es muy activo en nuestro país, que cuenta con grupos de investigación de reconocido prestigio en la comunidad internacional que investigan en PLN en español y las lenguas cooficiales, además de en otras lenguas. La Sociedad Española para el Procesamiento del Lenguaje Natural[1] (SEPLN), fundada en 1983, aglutina al grueso de investigadores de PLN, edita una revista de carácter científico, organiza un congreso internacional al año y recopila toda la información en relación a los grupos de investigación y docencia en PLN que se realiza en el estado.

Ya se está reconociendo desde la Administración el papel fundamental que juegan las tecnologías del lenguaje en el desarrollo de la sociedad digital del futuro. El Plan Nacional de Impulso de las Tecnologías del Lenguaje[2], presentado en 2015, se ejecuta en el marco de la Agenda Digital para España y tiene como objetivo fomentar el desarrollo del PLN y la traducción automática en lengua española y lenguas cooficiales.

[1] http://www.sepln.org/

[2] https://www.plantl.gob.es/tecnologias-lenguaje/PTL/Paginas/plan-impulso-tecnologias-lenguaje.aspx

"Desde los años 50 se investiga para que los ordenadores sean capaces de procesar el lenguaje humano, comprender su significado y generar lenguaje."

Este plan tiene cinco ejes, pero hay dos de especial relevancia para el tema que nos ocupa. El Eje III presenta a la Administración Pública como impulsora de la industria del lenguaje jugando dos papeles diferentes: (i) beneficiaria de sus aplicaciones con el objetivo de ofrecer servicios públicos más innovadores e inteligentes al ciudadano, y (ii) generadora de recursos lingüísticos fundamentales para la investigación, debido al gran valor potencial que tiene buena parte de la información que genera el sector público para el desarrollo de la industria del PLN. Por otra parte, el Eje IV se centra en actuaciones en servicios públicos concretos de alto impacto social que den lugar a productos y servicios acabados. Se definen como sectores estratégicos: Sanidad, Turismo y Educación.

Recientemente se ha presentado la Estrategia española de I+D+i en Inteligencia Artificial[3] en la que las Tecnologías del Lenguaje aparecen como una de las tecnologías principales en todas las áreas estratégicas priorizadas. En lo que respecta a la IA para la sociedad estas áreas son: Administración pública, Educación, Ciudades y Territorios Inteligentes, y Salud.

En este artículo entraremos en más detalle en las tecnologías del lenguaje aplicadas al sector estratégico de la sanidad.

2. TECNOLOGÍAS DEL LENGUAJE EN EL SECTOR DE LA SANIDAD

a. Información en el ámbito sanitario

Buena parte de la información en el ámbito sanitario son datos no estructurados, es decir, texto, que generalmente se puede clasificar en tres grandes tipos:

- Literatura científica disponible en bibliotecas digitales o portales específicos. Se trata de documentos con estructura cuya redacción se puede considerar correcta tanto desde el punto de vista estilístico, como sintáctico y ortográfico.

- Historia clínica electrónica (HCE), que recoge la información sobre la salud de un paciente (diagnósticos, medicamentos, pruebas de diversa índole, alergias, vacunaciones, planes de tratamiento). Este texto libre puede incluir errores ortográficos, estructuras sintácticas incorrectas, uso de jerga, sinónimos, acrónimos y abreviaturas sin su forma extendida.

- Redes sociales, comunidades online de pacientes y documentación explicativa a pacientes. Al igual que con la HCE, y especialmente en los blogs y micromensajes, el lenguaje utilizado puede alejarse de la norma, presentar incorrecciones, además de utilizar símbolos específicos de las propias redes (emoticonos, hashtags, etc.).

Toda esta información está sujeta a un crecimiento constante y el objetivo general de su procesamiento es obtener conocimiento no trivial y útil de ella. Cada tipo de información puede generar conocimiento de diversa índole y sus características específicas suponen diferentes niveles de dificultad en su procesamiento.

b. Procesamiento básico de textos biomédicos

Hay una serie de procesamientos que podríamos considerar básicos o generales para comprender el contenido del documento en cuestión y que serán necesarios para poder abordar con éxito casos de uso relevantes. Son los siguientes:

- Reconocimiento de conceptos o entidades biomédicos. Se trata de reconocer y clasificar por categorías de forma automática conceptos biomédicos en los textos, como enfermedades, medicamentos, etc. [1].

- Desambiguación de acrónimos y abreviaturas. El uso de acrónimos en textos clínicos está muy extendido y normalmente se refiere a un concepto o entidad que se utiliza con frecuencia, de ahí la necesidad de utilizar una forma abreviada. La ambigüedad intrínseca en muchos acrónimos de este dominio dificulta encontrar a qué entidad concreta se refiere [2].

- Identificación de información temporal relacionada a eventos. Se trata de identificar el marco temporal asociado a síntomas, diagnósticos, etc. [3].

[3] http://www.ciencia.gob.es/portal/site/MICINN/menuitem.26172fcf4eb029fa6ec7da6901432ea0/?vgnextoid=70fcdb77ec929610VgnVCM1000001d04140aRCRD

- Identificación de modificadores como negación, presencia/ausencia y especulación, que facilitan la comprensión correcta del texto [4].

3. ALGUNOS CASOS DE USO EN EL SECTOR DE LA SANIDAD

Apoyo a la toma de decisiones clínicas

Los sistemas de ayuda a la decisión clínica son sistemas de información diseñados para mejorar la toma de decisiones. Estos sistemas hacen sugerencias para que el clínico las revise, elija la información útil y descarte sugerencias erróneas. El papel del PLN en este caso de uso es obtener conocimiento directamente de la HCE, que se podrá combinar con datos cuantitativos de pruebas realizadas, conocimiento previamente obtenido de otros casos similares y de literatura científica. En [5] se presenta un resumen de las técnicas de PLN que se utilizan en este tipo de sistemas.

Asistencia en la codificación de HCE (CIE-10)

La codificación médica según la Clasificación Internacional de Enfermedades (CIE) consiste en la asignación de códigos estandarizados a HCE en representación de diagnósticos y procedimientos, cuya finalidad es la generación de estadísticas de morbilidad y mortalidad. La nueva versión CIE-10-ES como sistema de codificación clínica es obligatoria en los centros sanitarios españoles a partir del 01/01/2016 y consta de casi 70.000 diagnósticos y unos 72.000 procedimientos. Para los centros sanitarios llevar a cabo esta codificación tiene un coste enorme ya que se realiza de forma manual (o mínimamente asistida). Cuando se trata de la codificación automática de informes cortos, como partes de defunción o informes estructurados con un diagnóstico o procedimiento conciso por campo, los sistemas automáticos propuestos en el estado del arte ofrecen resultados muy competitivos [6]. Sin embargo, la mayor parte de la HCE en nuestro sistema sanitario son documentos con una estructura muy limitada y no necesariamente cortos, por lo que se trata fundamentalmente de texto libre y los sistemas existentes actualmente llegan a asistir a los codificadores, pero sus resultados no se pueden dar por correctos sin una validación humana [7].

"Los sistemas de ayuda a la decisión clínica son sistemas de información diseñados para mejorar la toma de decisiones. Estos sistemas hacen sugerencias para que el clínico las revise, elija la información útil y descarte sugerencias erróneas."

Extracción de relaciones

La identificación de relaciones significativas entre las entidades biomédicas y sucesos identificados en los textos médicos proporciona un conocimiento más profundo del contenido del texto que las propias entidades por separado. De ahí, que sea fundamental para la comprensión de un documento clínico poder establecer diferentes tipos de relaciones. Por ejemplo, un tipo de relación sería la detección de efectos adversos a medicamentos a partir de textos biomédicos. En [8] detectan efectos adversos en textos en español extraídos de las redes sociales.

Fenotipado computacional

Consiste en caracterizar grupos de población a partir de la HCE y para ello se utilizan técnicas de PLN. Es una tarea esencial porque tiene aplicaciones variadas, entre otras, la categorización de diagnósticos, el descubrimiento de nuevos fenotipos, el cribado de ensayos clínicos, la detección de eventos adversos de fármacos, etc. En [9] se puede encontrar una revisión sobre los avances recientes en sistemas de fenotipado computacional utilizando PLN.

Simplificación de textos para el paciente

Los textos médicos pueden ser difíciles de entender para los pacientes ya que están dirigidos a profesionales altamente cualificados y utilizan un lenguaje complejo y términos específicos para cada área. En 2018, un informe de la Organización Mundial de la Salud [10] concluye que la mayoría de los ciudadanos europeos no tiene suficientes conocimientos sobre la salud en lo que respecta a la capacidad de leer y comprender la información sanitaria, tomar decisiones adecuadas y seguir instrucciones en relación a su salud. El objetivo de la simplificación de textos consiste en la sustitución de dicha terminología por otra más fácil de entender para el paciente [11], y la sustitución de estructuras sintácticas complejas por otras más simples y fáciles de interpretar. Este último punto es el más complejo ya que hay que disponer de suficientes ejemplos de simplificación de textos biomédicos para poder aplicar las técnicas de aprendizaje supervisado.

"El español, es una de las tres lenguas más habladas en el mundo, junto con el chino y el inglés, de ahí que este sector industrial en España tenga unas perspectivas muy prometedoras si todos los agentes involucrados unen esfuerzos para su desarrollo y consolidación."

Escribas virtuales ambientales

Desde hace años el reconocimiento de voz se ha empleado en medicina por los sistemas de dictado automático para la transcripción de informes médicos [12]. En la actualidad, gracias a los sistemas de reconocimiento de voz los médicos obtienen la transcripción incluso en tiempo real y con un alto grado de efectividad y precisión. Dado que pueden surgir errores en el proceso que afecten al significado de algunas partes de la transcripción, se hacen estudios de los sistemas existentes para seguir mejorando su precisión [13].

4. OTROS CASOS DE USO EN LA ADMINISTRACIÓN

En general, los avances en traducción automática redundarán en una mejora de los servicios que requieran la traducción entre el castellano, las lenguas cooficiales y las lenguas comunitarias. Así mismo, las tecnologías del lenguaje mejorarán la accesibilidad a los servicios digitales de la Administración a personas con discapacidad: personas con discapacidad visual podrán interactuar a través de un asistente conversacional *(chatbot)*; además, actualmente existen algoritmos para reconocer el lenguaje de signos en tiempo real, cuyo resultado puede ser leído por un *chatbot,* posibilitando así que las personas que se comunican con el lenguaje de señas puedan acceder fácilmente a los servicios. Los sistemas de búsqueda de respuesta pueden ser de gran ayuda al ciudadano ya que éste puede realizar una consulta en lenguaje natural y el sistema le proporcionará la respuesta.

La capacidad de procesar y obtener conocimiento de grandes volúmenes de datos no estructurados relativos a convocatorias y contratación pública de todo tipo, análisis de oferta y demanda del mercado, noticias, publicaciones científicas, de divulgación, patentes etc., podrán ayudar a un mejor conocimiento del impacto de las políticas públicas.

En el sector estratégico del turismo los avances de la IA y el PLN tienen gran proyección, ya que es vital entender al cliente, sus necesidades, gustos y preferencias. Es muy común que se analicen las redes sociales, donde los usuarios de servicios como hoteles, restaurantes, museos o simplemente usuarios que viajan expresan sus opiniones acerca de sus vivencias en los diferentes lugares. Los sistemas de análisis automático de opiniones permiten extraer polaridad de sentimientos y emociones, lo que ayuda a la toma de decisiones de todos los agentes implicados [14].

En lo que respecta al área estratégica de la educación, el uso de PLN puede apoyar diferentes dominios de aprendizaje, como escritura, conversación, lectura, ayudando además a estudiantes con dificultades cognitivas. En [15] se da una visión general de los avances de PLN en el contexto de la educación, centrándose en las oportunidades y retos.

5. CONCLUSIONES

Las tecnologías del lenguaje humano no solo son esenciales para facilitar y mejorar todos los procedimientos relacionados con la comprensión y generación del lenguaje, sino que además constituye un sector industrial innovador. El español, es una de las tres lenguas más habladas en el mundo, junto con el chino y el inglés, de ahí que este sector industrial en España tenga unas perspectivas muy prometedoras si todos los agentes involucrados (empresas tecnológicas, investigadores, Administración) unen esfuerzos para su desarrollo y consolidación.

Los recientes avances de la IA en aprendizaje automático, especialmente en aprendizaje profundo *(deep learning),* ponen de manifiesto la importancia de disponer de grandes colecciones de datos que sean representativos del lenguaje natural y de los dominios concretos de aplicación. La Administración es la responsable de una buena parte de las colecciones de datos en sectores estratégicos, como sanidad, justicia o turismo, con lo que la disponibilidad de dichas colecciones por parte de los investigadores y las empresas tecnológicas redundaría finalmente en una mejora sustancial de la I+D+i. Para hacer posible esta disponibilidad, en algunos casos hay que poner el énfasis en definir métodos fiables y certificados de anonimización de documentos que salvaguarden la privacidad de los posibles ciudadanos implicados, contemplada en la ley de protección de datos. Pero este requisito debe ser superado lo antes posible y los agentes implicados de la Administración tomar conciencia de la importancia de facilitar la disponibilidad de dicha información. ✱

Bibliografía

[1] K. Gojenola, M. Oronoz, A. Perez, y A. Casillas, "IxaMed: Applying Freeling and a Perceptron Sequential Tagger at the Shared Task on Analyzing Clinical Texts", en *Proceedings of the 8th International Workshop on Semantic Evaluation*, 2014, pp. 361-365.

[2] I. Rubio-López, R. Costumero, H. Ambit, C. Gonzalo-Martín, E. Menasalvas, y A. Rodríguez. "Acronym Disambiguation in Spanish Electronic Health Narratives Using Machine Learning Techniques", *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, Vol. 235, pp. 251--255. 2017. DOI: 10.3233/978-1-61499-753-5-251.

[3] H. Lee, Y. Zhang, M. Jiang, J. Xu, C. Tao, y H. Xu. "Identifying direct temporal relations between time and events from clinical notes", *BMC Medical Informatics and Decision Making*, Vol. 18(2)", 2018.

[4] N.P. Cruz Díaz, y M.J. Maña López. *Negation and Speculation Detection*. Vol. 13. John Benjamins Publishing Company, 2019.

[5] J. A. Reyes-Ortiz, B. A. González-Beltrán y L. Gallardo-López, "Clinical Decision Support Systems: A Survey of NLP-Based Approaches from Unstructured Data," *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, Valencia, 2015, pp. 163-167.doi: 10.1109/DEXA.2015.47.

[6] M. Almagro, R. Martínez, S., V. Fresno. "A cross-lingual approach to automatic ICD-10 coding of death certificates by exploring machine translation", *Journal of Biomedical Informatics*, vol. 94, 2019.

[7] J. Pérez, A. Pérez, A. Casillas y K. Gojenola. "Cardiology record multi-label classification using latent Dirichlet allocation", *Computer Methods and Programs in Biomedicine*, Vol. 164, pp. 111--119, 2018.

[8] I. Segura-Bedmar, P. Martínez, R. Revert, J. Moreno-Schneider. "Exploring Spanish health social media for detecting drug effects", *BMC Medical Informatics and Decision Making*, 15 (2), 1–9, 2015.

[9] Z. Zeng, Y. Deng, X. Li, T. Naumann y Y. Luo. "Natural Language Processing for EHR-Based Computational Phenotyping". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 16(1), 2018, 139 – 153.

[10] World Health Organization "Health literacy", *The solid facts*, 2018

[11] J. Chen, E. Druhl, B. Polepalli, R. Thomas, K. Houston, C. ynthia, A. Brandt,D. MZulman, V. G. Vimalananda, S. Malkani, y H. Yu. "A Natural Language Processing System That Links Medical Terms in Electronic Health Record Notest to Lay Definitions: System Development Using Physician Reviews", *Journal of Medical Internet Research*, Vol. 20(1), 2018.

[12] H. Suominen, L. Zhou, L. Hanlen, y G. Ferraro, "Benchmarking Clinical Speech Recognition and Information Extraction: New Data, Methods, and Evaluations". *JMIR Medical Informatics*, 3(2), 2015.

[13] L. Zhou, S.V. Blackley, L. Kowalski, R. Doan, W.W Acker, A.B. Landman, E. Kontrient, D. Mack, M. Meteer, D.W. Bates y F.R. Goss. "Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists." *JAMA Network Open*, Vol. 1(3), 2018.

[14] A. Alaei, S. Becken y B. Stantic. "Sentiment Analysis in Tourism: Capitalizing on Big Data". *Journal of Travel Research*, Vol. 58(9), 2017.

[15] D. Litman. "Natural Language Processing for Enhancing Teaching and Learning". *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 4170-4176, 2016.