Position Paper: Tecnologías del Lenguaje aplicadas al sector de salud

Julio 2021

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

Contenido

Contextualización

La adopción de las Historias Clínicas Digitales por parte del sector sanitario está propiciando cambios muy significativos en la calidad de los servicios prestados. Además de usarse para la práctica clínica, los datos que contienen tienen el potencial de mejorar los servicios prestados mediante la generación de conocimiento y su aplicación a través de técnicas informáticas y analíticas. Entre los principales objetivos, se pueden incluir tanto el perfeccionamiento de las prácticas utilizadas y su aplicación, como el descubrimiento, análisis y evaluación de nuevos conocimientos que permitan mejorar la atención sanitaria.

Los datos se representan en formato estructurado y no estructurado. Gran parte de la información clínica se encuentra recogida en las notas médicas en formato de texto libre y por lo tanto, no estructurado, lo que supone un importante reto a la hora de extraer todo su valor. Dichas notas incluyen información sobre las condiciones del paciente, intervenciones realizadas, evolución y reacción frente a distintos tratamientos o su evolución clínica. De hecho, se estima que en torno a un 80% de la información clínica se encuentra almacenada en textos libres[1].

Sin embargo, a diferencia del contenido numérico, no es posible procesar directamente el texto mediante las herramientas de análisis que ayudan a los profesionales a realizar su trabajo de forma más eficiente. El tratamiento de este tipo de contenido de manera manual requeriría destinar muchos recursos sólo para esa tarea. Es en este punto, donde cobra especial importancia el Procesamiento del Lenguaje Natural (NLP).

Este conjunto de tecnologías permite aprovechar los textos médicos para extraer la información relevante para diversas aplicaciones. La capacidad de obtener datos discretos y procesables a partir de textos posibilita avanzar hacia un desarrollo del sector basado en el poder de los datos. Para ello, un elemento clave será la capacidad de integración de estas técnicas y herramientas en los flujos de trabajo de los usuarios sin cargar el peso sobre éstos, de forma que se promueva su adopción y uso a largo plazo.

Gracias al NLP, los ordenadores son capaces actualmente de analizar muchos más datos basados en texto y voz que las personas, de forma más rápida, consistente y objetiva. Por ello, considerando la gran cantidad de datos no estructurados que se generan cada día en los entornos sanitarios, desde registros médicos hasta estudios y resultados de investigaciones, la automatización de su procesamiento resulta decisiva para el continuo desarrollo y mejora de la calidad asistencial y la evolución hacia una asistencia clínica más personalizada.

Tecnologías relacionadas

Antes de explicar los diferentes casos de uso de las tecnologías de NLP en la salud, se van a definir las principales tecnologías relacionadas.

Modelos comunes de datos

Los datos sanitarios pueden variar mucho de una organización a otra en función de la finalidad de su registro (investigación clínica, trato directo con el paciente), los diferentes formatos

[1] Meystre S.M., Savova G.K., Kipper-Schuler K.C., Hurdle J.F. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 128–44 (2008)

utilizados por los sistemas de almacenamiento y los modelos de datos. Incluso un mismo concepto aún puede representarse de distintas maneras de una organización a otra a pesar de la creciente tendencia a la uniformización de vocabularios. En este contexto, es fundamental la estandarización de datos del sector sanitario para conseguir su interoperabilidad como un elemento crítico para el desarrollo del sector en el contexto actual.

Esta estandarización de datos permite transformarlos a un formato común para su análisis a gran escala, investigaciones colaborativas, o el desarrollo e intercambio de herramientas y metodologías universales. Cuantos más datos armonizados bajo un mismo estándar se compartan, mayores serán las posibilidades que se utilicen en la investigación de fármacos y terapias que puedan combatir enfermedades aun sin tratamientos eficaces.

Para ello, es necesario emplear un Vocabulario Estándar que permita representar los conceptos sin ambigüedad independientemente del lector o usuario. Algunos sistemas de codificación reconocidos son SNOMED CT[2] (*Systematized Nomenclature of Medicine - Clinical Terms*), centrado en terminología clínica y que permite una representación consistente del contenido de las historias clínicas electrónicas; RxNorm[3], que proporciona nombres normalizados para los medicamentos y que además permite vincularlos a otros vocabularios de medicamentos utilizados habitualmente; o CIE-11[4], undécima y última revisión de una lista generada por la OMS que contiene códigos para enfermedades, signos, síntomas, hallazgos anormales, quejas, circunstancias sociales y causas externas.

También es necesario establecer un Modelo Común de Datos (CDM) que sea capaz de incluir todos los datos del sector, desde historias clínicas hasta reclamaciones, o cualquier registro sanitario disponible, independientemente del modelo de almacenamiento de la fuente original. El CDM debe ser lo suficientemente exhaustivo como para poder representar la información en un amplio nivel de detalle, sin pérdida de valor, pero también comprensible para poder rellenarlo y consumirlo, aunque su característica más importante debe ser su amplia adopción.

Algunos ejemplos son el i2b2[5] (*Informatics for Integrating Biology and the Bedside*), que fue desarrollado hace más de una década y continúa siendo bastante popular, llegando a usarse en cientos de centros en todo el mundo y en redes a gran escala; y OMOP[6] (*Observational Medical Outcomes Partnership),* que se desarrolló como un modelo analítico compartido cuyo objetivo era estudiar los efectos de productos médicos en los pacientes. Dicho CDM tiene en cuenta la protección de datos, la estandarización al reutilizar vocabularios ya existentes y la escalabilidad, siendo uno de los CDM más utilizados a nivel global.

Ontologías

Las ontologías son especificaciones formales de los conceptos de un dominio y de las relaciones entre ellos, permitiendo modelar el conocimiento de un dominio en particular. Se pueden utilizar para anotar texto al trazar extractos de los textos con conceptos ontológicos que expresan el mismo significado, de modo que se puede dotar a dicho texto de cierta estructura semántica. Una vez que los textos están anotados, pueden utilizarse los conceptos identificados para

[2] https://www.snomed.org/
[3] https://www.nlm.nih.gov/research/umls/rxnorm/index.html
[4] https://icd.who.int/en
[5] https://www.i2b2.org/
[6] https://www.ohdsi.org/data-standardization/the-common-data-model/

realizar búsquedas sobre ellos, realizar análisis o clasificar pacientes[7]. En el ámbito médico, SNOMED CT[2] o la Ontología del Fenotipo Humano[8] (HPO) son ejemplos de ontologías ampliamente extendidas para la anotación de textos clínicos.

Interacción por voz

Las herramientas de reconocimiento de voz están basadas en técnicas de NLP y se han consolidado durante los últimos años. Los asistentes de voz y los programas de transcripción en tiempo real permiten ahorrar en tiempo y costes, a la vez que mejoran la experiencia de pacientes y profesionales, ya que permiten expresarse de una forma más directa, natural e intuitiva, y también llegar a personas que visualmente no puedan desenvolverse bien.

Los profesionales pueden utilizar esta tecnología para dictar notas a añadir en el sistema o actualizar la historia clínica del paciente. Menos tiempo escribiendo puede implicar más tiempo de trato directo con el paciente, mejorando así también la experiencia de éste. Por otra parte, combinar los asistentes conversacionales con esta tecnología, ofreciendo así la posibilidad de utilizar la voz como medio de comunicación, facilita la interacción y favorece la implicación del usuario.

Asistentes conversacionales

Los asistentes conversacionales se encargan de simular una conversación con una persona mediante respuestas automatizadas generadas por algoritmos en tiempo real. Se trata de una tecnología que ya ha irrumpido con fuerza en prácticamente todos los ámbitos, y el sanitario no es una excepción.

La asistencia médica encuentra en los asistentes conversacionales una nueva vía de comunicación bidireccional entre pacientes y profesionales, siendo la telemedicina o la Atención no Presencial uno de los campos que utilizan esta tecnología para el diagnóstico, la monitorización y el tratamiento de pacientes a distancia, con el objetivo de mejorar la salud de la población pudiendo llegar a más personas y prestando unos servicios más personalizados.

Estos asistentes no sustituyen la consulta médica con el profesional, sino que proporcionan mayor número de recursos a las personas para que puedan manejar situaciones de salud con más información. Algunos posibles usos incluyen la resolución de dudas sobre un tratamiento, la petición de cita, la valoración y el seguimiento del estado de ánimo del paciente, o el fomento y seguimiento de la adherencia a un tratamiento.

Su uso ofrece numerosas ventajas, entre las que se destacan que la comunicación se realice a distancia, sin necesidad de desplazamientos; la disponibilidad 24/7 y la inmediatez, pudiendo obtener respuestas en todo momento y al momento; o la posibilidad de preservar el anonimato, un tema de preocupación para aquellos que pueden padecer enfermedades asociadas a estigmas sociales, como las enfermedades mentales.

La crisis de la COVID-19 ha favorecido la aparición de nuevas iniciativas en el uso de bots sanitarios ante la dificultad de ofrecer una atención más convencional. Algunos ejemplos son el chatbot de Whatsapp de la OMS[9], cuyo objetivo es ofrecer datos y recomendaciones sobre la

[7] Kersloot M.G., van Putten F.J.P., Abu-Hanna A. et al. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. J Biomed Semant 11, 14 (2020)
[8] https://hpo.jax.org/app/
[9] https://www.who.int/news-room/feature-stories/detail/who-health-alert-brings-covid-19-facts-to-billions-via-whatsapp

situación actual; el chatbot de Whatsapp del Gobierno de España[10], en línea con el anterior, que busca ofrecer datos actualizados sobre la pandemia a nivel nacional y medidas para contener los contagios; o IMPAI[11], un chatbot desarrollado por un internista de la Sociedad Española de Medicina Interna que facilita el diagnóstico de COVID-19 y orienta sobre la fase o etapa de transmisibilidad en la que se encuentra el potencial paciente.

[10] https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/asuntos-economicos/Paginas/2020/080420-consulta.aspx
[11] https://www.fesemi.org/informacion/prensa/semi/impai-un-chatbot-que-impulsa-un-internista-de-semi-facilita-el-diagnostico-y

Casos de uso

A continuación, se exponen algunos casos de uso que ponen de manifiesto la importancia del NLP en el ámbito sanitario y las oportunidades que ofrecen para su desarrollo e innovación.

Análisis de patologías a partir de historias clínicas

El objetivo de este caso de uso es un mayor conocimiento sobre una patología concreta, incluyendo su curso evolutivo, su comportamiento, el tratamiento más adecuado en función del paciente, la identificación de los criterios de riesgo, la predicción de supervivencia, el control de síntomas derivados de la enfermedad, o la optimización de la gestión clínica. Para ello, es necesario estructurar los datos de los pacientes con dicha patología para facilitar el análisis y explotación de dicha información clínica.

A la hora de definir dicha estructuración, es necesario identificar su propósito para construir una ontología de dominio que incluya todos los aspectos que se quieren identificar (tratamiento, estadio de la enfermedad). Es posible apoyarse en ontologías predefinidas para la identificación de conceptos más generales como los síntomas. También es posible identificar una serie de patrones de interés que permitan extraer valores de los textos, como el número de plaquetas en sangre o la temperatura corporal de un paciente.

La aplicación de técnicas de NLP permitirá anotar los textos en función de su propósito, de modo que se estructure el conocimiento relevante del dominio para un uso sencillo que puede incluir herramientas de visualización que ayuden a reconstruir la cronología de la enfermedad y permitan un análisis más intuitivo.

Diagnóstico y pronóstico temprano

Una identificación correcta y anticipada de los síntomas es clave para un diagnóstico temprano. En este sentido, las técnicas de NLP apoyan al personal sanitario y a los sistemas de ayuda al diagnóstico mediante la recuperación automática de los síntomas identificados en las diversas notas asistenciales. De esta forma, también es posible identificar a otros pacientes con síntomas y tratamientos similares mediante técnicas de NLP.

Por estos motivos, las tecnologías de NP son clave en el diagnóstico temprano de enfermedades complejas, como las enfermedades raras, o en la búsqueda de comorbilidades de los pacientes crónicos y multipatológicos. Adicionalmente, la identificación automática de síntomas también permite relacionar los síntomas de un paciente con documentos científicos que facilitarían el diagnóstico de estas enfermedades difíciles de diagnosticar.

Detección de brotes y crisis sanitarias

La aplicación de algoritmos de análisis sobre textos médicos estructurados permitirá anticiparse a brotes epidémicos (enfermedades infecciosas, gripe o gastroenteritis, malaria, ébola, etc) de forma automática, facilitando la obtención de síntomas, síndromes y otros conceptos relevantes a partir de las notas médicas. De esta forma, se complementa y mejora la vigilancia de los departamentos de Salud Pública.

Las ventajas que ofrecen este tipo de soluciones apoyadas en NLP para la extracción automática de información se han comprobado recientemente, cuando algunas de estas herramientas12 dieron una primera alerta mundial sobre la crisis del COVID-19 al detectar un inusual aumento de casos de neumonía en Wuhan.

Ayuda en la salud mental

El estigma social y el desconocimiento que aún hoy rodea a las enfermedades mentales hace que muchas personas con trastornos mentales eviten buscar la ayuda de profesionales. Es aquí donde los chatbots pueden servir para ayudar a estas personas.

Los chatbots orientados a la salud mental permiten a las personas desahogarse sin sentir la vergüenza o la ansiedad de ser juzgados. Además, el servicio está siempre disponible, por lo que se puede usar a cualquier hora que se necesite. Los chatbots pueden aprender del usuario y adaptarse para ofrecer una orientación más personalizada. El análisis de sentimientos y la identificación de palabras o patrones clave también permitirán reconocer situaciones de riesgo, como el riesgo de suicidio, y alertar a profesionales.

Es importante remarcar que estos asistentes sirven para orientar al paciente y facilitar que se desahogue, pero no sustituyen a un profesional en ningún caso sustituyen a un profesional, aunque puedan realizar estas tareas complementarias.

[12] https://www.cnbc.com/2020/03/03/bluedot-used-artificial-intelligence-to-predict-coronavirus-spread.html

Retos del NLP en salud

La aplicación de las tecnologías de NLP en el campo de la salud debe abordar un conjunto de retos relacionados con aspectos de privacidad y seguridad de los datos sensibles de los ciudadanos, así como la disponibilidad de datos anotados para un correcto entrenamiento de los modelos.

Anonimización

La anonimización es un elemento fundamental para la interoperabilidad de datos sanitarios y la información generada, ya que dichos datos están catalogados como de máxima sensibilidad. La aplicación de estas técnicas permitirá una compartición legal y segura de datos entre organizaciones, de modo que se puedan utilizar estos datos para extraer su máximo valor sin comprometer la integridad y privacidad de las personas.

En este escenario, las técnicas de anonimización de NLP tienen un papel protagonista al minimizar el riesgo de re-identificación de los individuos a partir de los datos, ya sean en formato texto o voz, eliminando toda referencia a su identidad, pero conservando intacta la veracidad de la información. Así, la anonimización debe adoptarse como proceso fundamental que habilite el tratamiento seguro de datos de tipo personal.

Resultan de particular interés los métodos basados en categorías (nombre, organización, etc.) asociadas a cada unidad de información que posteriormente pueden ser eliminadas o sustituidas. No obstante, es importante invertir en estos casos en técnicas de NLP que permitan identificar apropiadamente el contexto, así como posibles errores ortográficos, de modo que se obtengan soluciones robustas adaptadas a las particularidades del lenguaje natural.

Estandarización

Tal y como se ha mencionado previamente, tanto los datos sanitarios como su formato o modelo de almacenamiento pueden variar mucho entre organizaciones. El proceso de estandarización permite transformarlos a un formato común, de modo que puedan realizarse análisis e investigaciones sobre fuentes de datos heterogéneas. Por ello, junto con la anonimización, la estandarización es un elemento crucial en la interoperabilidad de datos, por lo que ambas técnicas resultan críticas para el desarrollo del sector en el contexto actual.

Disponibilidad de datos anotados

Al contrario que en textos de carácter más general, la narrativa clínica suele estar escrita por y dirigida a profesionales de la salud. Por lo tanto, se utiliza mucha jerga médica, lo que implica que es necesario utilizar corpus médicos para que las máquinas puedan aprender este lenguaje. En este sentido, las distintas especialidades médicas añaden una capa extra de complejidad, puesto que, por ejemplo, un dermatólogo utilizará términos muy distintos a los que pueda emplear un cirujano. Por ello, para una mayor precisión, los algoritmos se deberán entrenar con datos distintos en función del ámbito y especialidades médicas que los utilicen.

Además, dichos datos se podrán referir a los motivos de una consulta, resultados de pruebas médicas, tratamientos suministrados, registros de estancia en un hospital o cualquier otra información relevante de cara a futuras tomas de decisiones. La diferencia de propósitos hace que los datos sean muy heterogéneos en cuanto a contenido y nivel de detalle[13]. Por ejemplo,

[13] Leaman R., Khare R., Lu Z. Challenges in clinical natural language processing for automated disorder normalization. Journal of Biomedical Informatics 57: 28-37 (2015)

mientras que la narrativa de una publicación tiene una estructura sintáctica más clara, una nota médica a menudo se rellena a contrarreloj, por lo que es habitual encontrar frases esquemáticas, omisión de información que puede inferirse del contexto, acrónimos e incluso errores ortográficos.

Los corpus narrativos clínicos disponibles para su uso son relativamente escasos por motivos de privacidad, y la mayor parte de ellos están en inglés. Estos corpus son necesarios para entrenar los algoritmos, por lo que la falta de disponibilidad puede entorpecer el desarrollo del sector. Los avances y la inversión en técnicas de anonimización son la clave para abordar este reto.

Situación en España

El pasado año el Gobierno de España presentó el Plan de Recuperación, Transformación y Resiliencia[14], documento en el que se esbozan las inversiones a realizar durante los próximos años para la recuperación económica del país. Este plan pone énfasis en la renovación del sistema público de salud mediante la innovación tecnológica y métodos avanzados que mejoren el tratamiento de datos, destacando también la importancia de impulsar una Estrategia Nacional de Inteligencia Artificial[15] (ENIA) que permita liderar el desarrollo de estas tecnologías en español, entre las que se encuentra el NLP. Está claro que si bien gran parte de las tecnologías del lenguaje se desarrollan en inglés, otras lenguas como el español son muy representativas en el campo médico, por lo que es necesario generar recursos también en este idioma para poder beneficiarnos de los avances.

Ya en 2015 se anunció la puesta en marcha desde el Gobierno del Plan de Impulso de las Tecnologías del Lenguaje[16] (Plan TL), dotado de casi 90 millones de euros para los siguientes 5 años. Dentro del marco de este plan, se han financiado a lo largo de estos años numerosos eventos y competiciones, pero también el desarrollo de estudios, herramientas de software e infraestructuras para el ámbito sanitario y que se ofrecen disponibles para su uso.

[14] https://www.lamoncloa.gob.es/temas/fondos-recuperacion/Documents/30042021-Plan_Recuperacion_%20Transformacion_%20Resiliencia.pdf
[15] https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/ENIAResumen2B.pdf
[16] https://plantl.mineco.gob.es/Paginas/index.aspx