Cómo gestionar la sobrecarga de información científica sobre COVID-19

theconversation.com/como-gestionar-la-sobrecarga-de-informacion-cientifica-sobre-covid-19-138651

Arantxa Otegi, Aitor Soroa, Eneko Agirre, Jon Ander Campos

Desde el inicio de la crisis sanitaria provocada por la COVID-19, los científicos que luchan contra esta enfermedad están ahogados por la creciente literatura científica.

Ante esta situación, y respondiendo a un llamamiento del gobierno de los Estados Unidos, numerosos grupos de investigadores han explorado diferentes soluciones. El sistema de búsqueda de respuestas que hemos propuesto los autores de este artículo ha sido uno de los premiados por esa iniciativa.

Miles de artículos por semana

La comunidad médica y científica necesita compartir información relevante para hacer frente a la pandemia de COVID-19. Sin embargo, la cantidad de información disponible hoy día acerca del coronavirus causante de esta enfermedad es enorme.

Además, conforme pasa el tiempo y a medida que la pandemia se ha ido extendiendo por todo el mundo, el ritmo de publicación de artículos científicos sobre este tema ha ido creciendo.

Se han llegado a publicar más de 4 000 *papers* en una semana. Expertos como el virólogo Timothy Sheahan, que trabaja en la Universidad de Carolina del Norte, han reconocido la dificultad de estar al corriente de todo lo que se publica.

Llamamiento a los investigadores de IA

Ante esta situación, y a petición de la Oficina de Política de Ciencia y Tecnología de la Casa Blanca, varios grupos de investigación destacados pusieron a disposición de la comunidad científica mundial una colección de artículos científicos: COVID-19 Open Research Dataset (CORD-19), con más de 63 000 documentos.

Además, se hizo un llamamiento a los investigadores de todo el mundo para que aplicaran las últimas técnicas en inteligencia artificial y procesamiento del lenguaje. El objetivo era conseguir que los científicos que luchan contra la enfermedad COVID-19 puedan encontrar información relevante y precisa en las publicaciones.

Los organizadores pusieron en marcha una competición a través de la plataforma Kaggle. En una primera fase se definieron 10 tareas. En cada una de ellas se enumeraron las preguntas clave de un tema diferente relacionado con la COVID-19. Estas preguntas fueron creadas basándose, entre otros, en el plan de acciones de investigación y desarrollo de la Organización Mundial de la Salud.

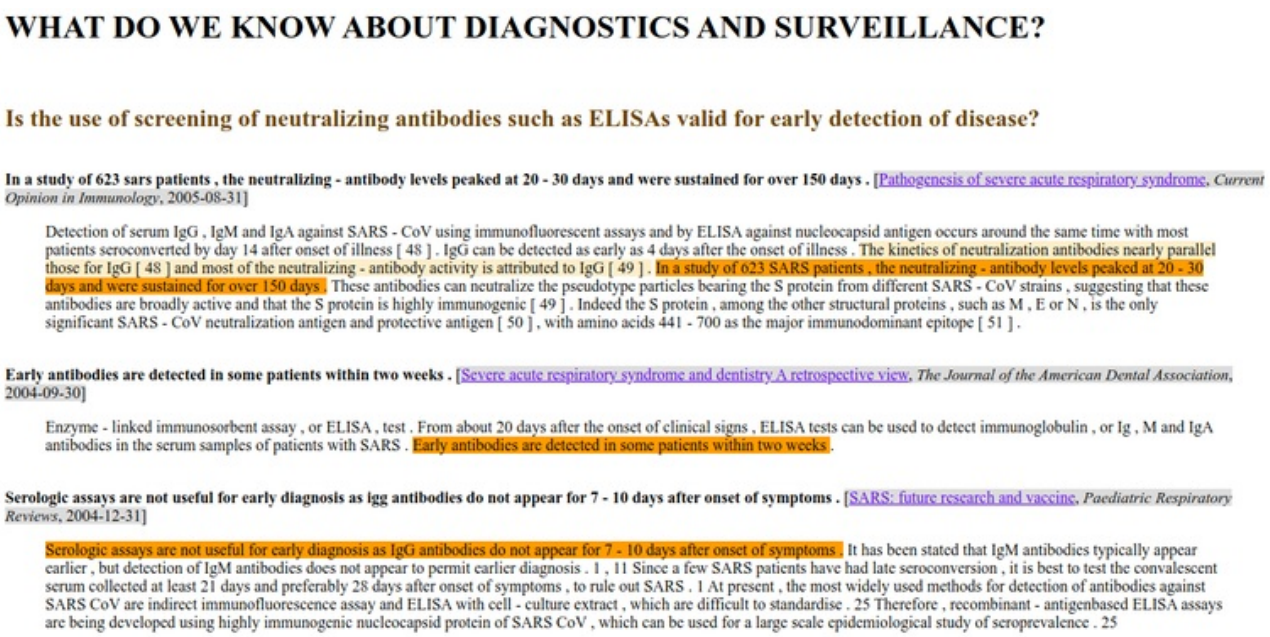
Los investigadores participantes han puesto en esta plataforma los sistemas de procesamiento de datos y texto desarrollados para esta competición, de manera que están disponibles para expertos de todo el mundo.

Un sistema que responde a las preguntas

El grupo de investigación Ixa participamos en esta competición. Para ello desarrollamos un sistema que, analizando los mencionados artículos científicos, busca respuestas a las preguntas planteadas por los expertos.

Nuestro sistema ganó una de las 10 tareas de la primera fase Concretamente, ha sido seleccionado como el sistema que mejor ha respondido al cuestionario sobre el tema *¿Qué sabemos sobre diagnóstico y vigilancia?*

En la imagen que sigue a este párrafo se puede observar una de las preguntas de este tema y lo que el sistema responde (en negrita), así como información de la publicación y contexto donde se ha encontrado la respuesta (en naranja oscuro la respuesta, en naranja más claro la información más relevante).



WHAT DO WE KNOW ABOUT DIAGNOSTICS AND SURVEILLANCE?

Is the use of screening of neutralizing antibodies such as ELISAs valid for early detection of disease?

In a study of 623 sars patients , the neutralizing - antibody levels peaked at 20 - 30 days and were sustained for over 150 days . [Pathogenesis of severe acute respiratory syndrome, *Current Opinion in Immunology*, 2005-08-31]

Detection of serum IgG , IgM and IgA against SARS - CoV using immunofluorescent assays and by ELISA against nucleocapsid antigen occurs around the same time with most patients seroconverted by day 14 after onset of illness [48] . IgG can be detected as early as 4 days after the onset of illness . The kinetics of neutralization antibodies nearly parallel those for IgG [48] and most of the neutralizing - antibody activity is attributed to IgG [49] . In a study of 623 SARS patients , the neutralizing - antibody levels peaked at 20 - 30 days and were sustained for over 150 days . These antibodies can neutralize the pseudotype particles bearing the S protein from different SARS - CoV strains , suggesting that these antibodies are broadly active and that the S protein is highly immunogenic [49] . Indeed the S protein , among the other structural proteins , such as M , E or N , is the only significant SARS - CoV neutralization antigen and protective antigen [50] , with amino acids 441 - 700 as the major immunodominant epitope [51] .

Early antibodies are detected in some patients within two weeks . [Severe acute respiratory syndrome and dentistry A retrospective view, *The Journal of the American Dental Association*, 2004-09-30]

Enzyme - linked immunosorbent assay , or ELISA , test . From about 20 days after the onset of clinical signs , ELISA tests can be used to detect immunoglobulin , or Ig , M and IgA antibodies in the serum samples of patients with SARS . Early antibodies are detected in some patients within two weeks .

Serologic assays are not useful for early diagnosis as igg antibodies do not appear for 7 - 10 days after onset of symptoms . [SARS: future research and vaccine, *Paediatric Respiratory Reviews*, 2004-12-31]

Serologic assays are not useful for early diagnosis as IgG antibodies do not appear for 7 - 10 days after onset of symptoms . It has been stated that IgM antibodies typically appear earlier , but detection of IgM antibodies does not appear to permit earlier diagnosis . 1 , 11 Since a few SARS patients have had late seroconversion , it is best to test the convalescent serum collected at least 21 days and preferably 28 days after onset of symptoms , to rule out SARS . 1 At present , the most widely used methods for detection of antibodies against SARS CoV are indirect immunofluorescence assay and ELISA with cell - culture extract , which are difficult to standardise . 25 Therefore , recombinant - antigenbased ELISA assays are being developed using highly immunogenic nucleocapsid protein of SARS CoV , which can be used for a large scale epidemiological study of seroprevalence . 25

Respuestas del sistema dadas a una de las preguntas del tema *What do we know about diagnostics and surveillance?*

Todas las preguntas y las respuestas dadas por el sistema pueden verse aquí y el código se puede consultar junto con su descripción técnica.

¿Cómo se realiza la búsqueda?

Ya hemos visto a qué tipo de preguntas responde este exitoso sistema de búsqueda de respuestas. Pero ¿cómo busca el sistema estas respuestas entre tantos artículos científicos? El proceso de búsqueda de respuestas para una pregunta concreta se divide en 3 fases principales.

- En una primera fase se seleccionan de toda la colección de artículos solamente los que están relacionados con la enfermedad COVID-19, ya que en esta colección también se incluyen artículos sobre otros coronavirus distintos al COVID-19, como SARS-CoV y MERS. Para realizar esta selección, se analiza el título y resumen de cada trabajo para ver si contienen palabras utilizadas como sinónimo de la COVID-19 por la comunidad científica.

- En la siguiente fase un sistema de recuperación de información extrae unos pocos artículos de entre los previamente seleccionados. El sistema es capaz de discriminar los artículos que potencialmente contienen la repuesta a la pregunta formulada por el usuario. Para ello, primero se crea una estructura de datos llamada índice que guarda una referencia del artículo donde aparece cada palabra. Esta estructura de datos permite buscar información de forma muy eficaz.

  Una vez creado el índice, se utiliza elalgoritmo de búsqueda BM25 para encontrar los artículos más relevantes para cada pregunta. Dicho algoritmo utiliza el índice para buscar en qué artículos se encuentra cada palabra de la pregunta. BM25 asigna una puntuación que mide la relevancia de cada uno de los artículos para cada pregunta. Para ello tiene en cuenta diferentes métricas como el número de apariciones y la longitud de los artículos. En esta fase se seleccionan los 20 artículos con mayor puntuación.

- En la fase final, la búsqueda de respuestas se hace sobre los 20 artículos seleccionados. Para ello se aplican técnicas avanzadas basadas en redes neuronales de inteligencia artificial. En concreto, estas técnicas emplean el modelo lingüístico denominado BERT (Bidirectional Encoder Representations from Transformers).

BERT, utilizado en el buscador de Google, es capaz de crear una representación contextual para cada palabra, que depende también de las que le rodean. Las palabras y expresiones que tienen un significado parecido estarán más cerca entre ellas que las que no lo tienen, como si de un mapa se tratara.

Para adaptar este modelo lingüístico y darle la capacidad de responder preguntas se utilizaron 83 000 preguntas y respuestas anotadas por humanos. Es importante puntualizar que estas 83 000 preguntas no tienen relación con la enfermedad y tratan sobre temas generales. Por ello, el sistema podría ser utilizado también para responder preguntas en otros dominios y en un futuro podría adaptarse mejor al tema.

Tras aplicar este último sistema de búsqueda de respuesta, el usuario que formula la pregunta recibe cinco artículos de los seleccionados en la segunda fase y en cada uno se resaltan las tres respuestas más probables.

Limitamos el número de artículos a cinco para no sobrecargar al usuario con demasiada información. Además, cabe la posibilidad de que no todos los 20 artículos de la segunda fase contengan la respuesta deseada y por ello también son descartados.

Un sistema útil para los científicos

Este sistema de búsqueda de respuestas es de gran ayuda para buscar de una manera rápida y cómoda respuestas a las preguntas prioritarias de los expertos sobre la COVID-19, disminuyendo el tiempo necesario para recabar información.

Por ejemplo, el programa puede responder a preguntas sobre la historia del coronavirus, la transmisión y diagnóstico, las medidas de prevención en el contacto entre seres humanos y animales y las lecciones de estudios epidemiológicos previos.

Los últimos avances en el uso de la inteligencia artificial para el procesamiento del lenguaje han permitido desarrollar sistemas avanzados de acceso a la información. En un trabajo relacionado de nuestro grupo hemos demostrado que pueden llegar a tener conversaciones sobre temas especializados.

Estos sistemas son prueba de la importancia de estas tecnologías para hacer frente a la gran cantidad de información que se genera constantemente.