

An automatic information extraction system for scientific articles on COVID-19

VIGICOVID is a system that uses natural language questions to get answers in the avalanche of information on COVID-19 and SARS-CoV-2

RESEARCH 23/03/2022



Eneko Agirre and Xabier Saralegi. Photo: UPV/EHU

Researchers from the UPV/EHU-University of the Basque Country, the UNED (National Distance Education University) and Elhuyar have created the VIGICOVID system, thanks to Supera COVID-19 (Overcoming COVID-19) funding by the CRUE (Association of Spanish Universities). This system addresses the need to search for answers in the avalanche of information generated by all the research conducted across the world relating to the pandemic. By means of artificial intelligence, the system displays the answers found in a set of scientific articles in an orderly fashion, and uses natural language questions and answers.

The global bio-health research community is making a tremendous effort to generate knowledge relating to COVID-19 and SARS-CoV-2. In practice, this effort means a huge, very rapid production of scientific publications, which makes it difficult to consult and analyse all the information. That is why experts and decision-making bodies need to be provided with information systems to enable them to acquire the knowledge they need.

This is precisely what has been explored in the VIGICOVID researchers project run by the UPV/EHU's HiTZ Centre, the UNED's NLP & IR group, and Elhuyar's Artificial Intelligence and Language Technologies Unit, thanks to Fondo Supera COVID-19 funding awarded by the CRUE. In the study, under the coordination of the UNED research group they have created a prototype to extract information through questions and answers in natural language from an updated set of

scientific articles on COVID-19 and SARS-CoV-2 published by the global research community.

"The information search paradigm is changing thanks to artificial intelligence," said Eneko Agirre, head of the UPV/EHU's HiTZ Centre. "Until now, when searching for information on the internet, a question is entered, and the answer has to be sought in the documents displayed by the system. However, in line with the new paradigm, systems that provide the answer directly without any need to read the whole document are becoming more and more widespread."

In this system, "the user does not request information using keywords, but asks a question directly", explained Elhuyar researcher Xabier Saralegi. The system searches for answers to this question in two steps: "Firstly, it retrieves documents that may contain the answer to the question asked by using a technology that combines keywords with direct questions. That is why we have explored neural architectures," added Dr Saralegi. Deep neural architectures fed with examples were used: "That means that search models and question answering models are trained by means of deep machine learning."

Once the set of documents has been extracted, they are reprocessed through a question and answer system in order to obtain specific answers: "We have built the engine that answers the questions; when the engine is given a question and a document, it is able to detect whether or not the answer is in the document, and if it is, it tells us exactly where it is," explained Dr Agirre.

A readily marketable prototype

The researchers are satisfied with the results of their research: "From the techniques and evaluations we analysed in our experiments, we took those that give the prototype the best results," said the Elhuyar researcher. A solid technological base has been established, and several scientific papers on the subject have been published. "We have come up with another way of running searches for whenever information is urgently needed, and this facilitates the information use process. On the research level, we have shown that the proposed technology works, and that the system provides good results," Agirre pointed out.

"Our result is a prototype of a basic research project. It is not a commercial product," stressed Saralegi. But such prototypes can be modelled easily within a short time, which means they can be marketed and made available to society. These researchers stress that artificial intelligence enables increasingly powerful tools to be made available for working with large document bases. "We are making very rapid progress in this area. And what is more, everything that is investigated can readily reach the market," concluded the UPV/EHU researcher.

bibliographic reference

Arantxa Otegi, Iñaki San Vicente, Xabier Saralegi, Anselmo Peñas, Borja Lozano, Eneko Agirre
Information retrieval and question answering: A case study on COVID-19 scientific literature
Knowledge-Based Systems
DOI (Digital Object Identifier): 10.1016/j.knosys.2021.108072

image gallery