National Institute of Informatics News ISSN 1883-1966 (Print) ISSN 1884-0817 (Online)

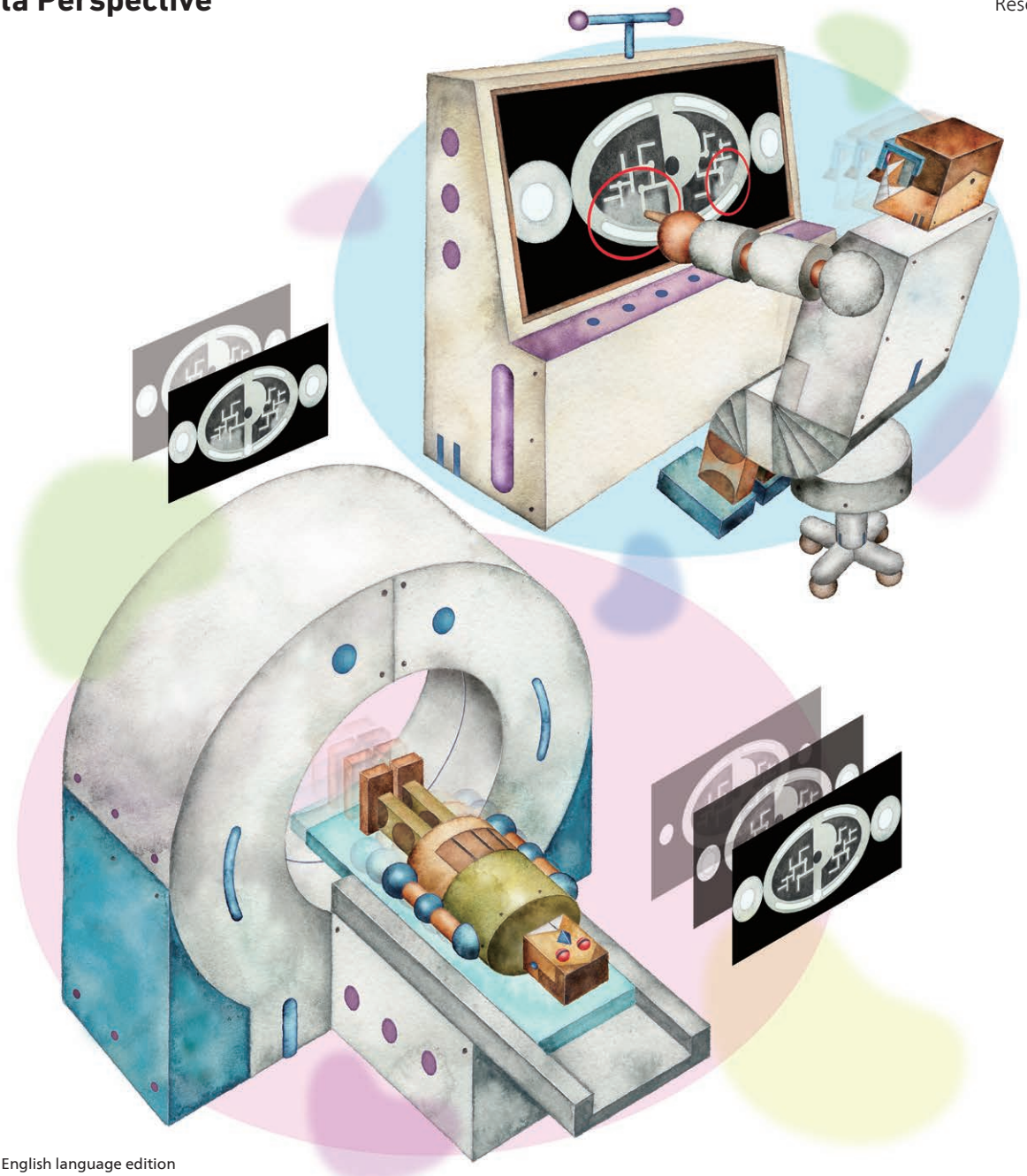
NII Today

75
Sep. 2020
National Institute of Informatics News

Feature

Confronting COVID-19 The Challenges of Informatics

A Look at COVID-19 From a Data Perspective



NII Interview

How CT Image Analysis is Helping to Combat Infectious Diseases

Joint development of AI with the academic community and universities to help in the diagnosis of COVID-19 pneumonia

AOKI Shigeki [President, Japan Radiological Society]
SATOH Shin'ichi [Professor, Digital Content and Media Sciences Research Division, NII / Director, Research Center for Medical Bigdata]

Building a Website to Offer a Global Overview of COVID-19 Information

A collaboration between natural language processing researchers

AIZAWA Akiko [Professor / Vice Director General, NII Digital Content and Media Sciences Research Division

Commentary **Research Through Open Collaboration in the Post-Covid Era**

KAWAHARA Daisuke [Professor, Department of Computer Science and Communications Engineering, School of Fundamental Science and Engineering, Waseda University

Exploring Social Media's "Fake News" Problem

The importance of questioning information sources

TORIUMI Fujio [Associate Professor, Department of Systems Innovation, Graduate School of Engineering, The University of Tokyo

Big Data Analysis for Resolving the Social Problems of COVID-19

Visualizing the "self-restraint" rate for going out and human contact frequency

MIZUNO Takayuki [Associate Professor, Information and Society Research Division, NII]

This English language edition NII Today corresponds to No. 89 of the Japanese edition

How CT Image Analysis is Helping to Combat Infectious Diseases

Joint development of AI with the academic community and universities to help in the diagnosis of COVID-19 pneumonia

AOKI Shigeki
Professor, Department of Radiology, Graduate School of Medicine, Juntendo University / President, Japan Radiological Society

SATO Shin'ichi
Professor, Digital Content and Media Sciences Research Division, NII / Director, Research Center for Medical Bigdata

Interviewer: TAINAKA Madoka
Senior Editor, NII Today / Science writer

COVID-19 is currently diagnosed using PCR tests, but for more precise diagnosis and prognosis, computed tomography (CT) is also a useful tool. Japan boasts more CT systems per capita in its healthcare facilities than any other country and CT medical imaging is seen as highly promising. NII's Research Center for Medical Bigdata (RCMB), together with the Japan Radiological Society (JRS), has been developing AI (artificial intelligence) for using collected CT images to help in the diagnosis of COVID-19 pneumonia. To find out more, we interviewed JRS president AOKI Shigeki and RCMB director SATOH Shin'ichi about the results of this effort, as well as RCMB initiatives, and the potential of AI for medical applications.

CT scans for more precise diagnosis and prognosis

—— How are CT scans used to diagnose COVID-19?

AOKI CT images of the lungs of people with COVID-19 often show ground glass opacity or reticular shadows characteristic of viral pneumonia. The opacity or shadows are typically found not just in one location but also at the periphery, and on both lungs. (See Fig. 1.) In the case of the large outbreak on the Diamond Princess cruise ship, for instance, 112 people who tested positive on PCR tests also underwent a CT scan. Of these, 30 were symptomatic and 82 were not. The characteristic shadows of COVID-19 were observed in the lung CT images of more than half of the 82 asymptomatic cases (44 cases). CT is much more sensitive than X-rays, so it is useful for detecting minute lesions.

—— It's surprising that over half of the asymptomatic cases showed lung inflammation.

AOKI The human body has a remarkable residual function capacity, so many times tissues continue to function within their normal range even after their shape has been altered. Take the liver, which is known as a "silent organ." Sometimes no symptoms will be evident even after it has shrunk and hardened dramatically due to cirrhosis. CT scans are effective for finding lesions in asymptomatic people and for ascertaining symptoms more precisely.

However, the JRS does not recommend CT for COVID-19 screening. The reason is that despite its high sensitivity, CT has a relatively low specificity. This means that there is a high likelihood of false positive results. It is difficult to distinguish COVID-19 pneumonia from other viral pneumonias, like those arising from pulmonary fibrosis. And because changes in the early stages of infection may be very slight, it is difficult to discover lesions without specialist expertise.

Mind you, if no other viral pneumonias are present, the use of CT scans for COVID-19 screening of patients who are free of lung disease can be effective. As the JRS guidelines subtly put it, CT testing is not recommended as a substitute for a PCR test, but it is acceptable for helping to make medical decisions with patients suspected of having COVID-19.

—— What kinds of people should get a CT scan?

AOKI Anyone who is symptomatic and tests positive on PCR or whose condition is worsening should get a CT scan. CT images are useful for examining lesions in detail and for making a prognosis, regardless of whether the patient is getting better or worse.

Using the diagnoses of radiologists as training data for improving AI precision

—— The JRS and RCMB are jointly developing AI to support the diagnosis of COVID-19. How did this initiative come

AOKI Shigeki

Born in Tokyo in 1959. Graduated from the Faculty of Medicine at the University of Tokyo in 1984. After clinical training, in 1987 he studied neuroradiology at UCSF (University of California, San Francisco). In 1995 he became an assistant professor in the Department of Radiology at the University of Yamanashi. In 2000 he became an assistant professor in the Department of Radiology at the University of Tokyo Graduate School of Medicine. He has been in his current position since 2008. He currently serves as president of the Japan Radiological Society and secretary general of the Japanese Society for Magnetic Resonance in Medicine. He is a member of the Technical Assessment Committee of the Central Social Insurance Medical Council. In 2012 he chaired the 107th National Examination for Medical Practitioners.

SATO Shin'ichi

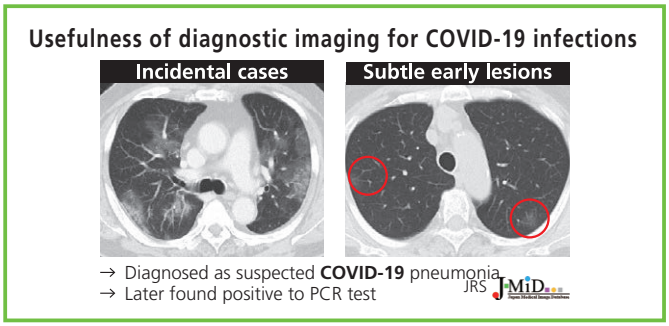


Fig. 1 The image at left shows a case of COVID-19 pneumonia without symptoms that was found incidentally. Infection was suspected based on the characteristic CT findings and later confirmed by a positive PCR test result. CT is useful in the diagnosis of COVID-19. However, in some cases CT is unable to indicate lesions and sometimes at the early stage of infection the changes are so slight that it is difficult for anyone without specialist expertise to observe lesions, as shown by the red circle in the photo at right. It is hoped that AI will make it easier to identify lesions like these, which are difficult for humans to notice. (Image courtesy of Associate Professor AKASHI Toshiaki, Department of Radiology, Juntendo University School of Medicine.)

about?

AOKI Around five years ago, the JRS started a project to collect all CT images from The University of Tokyo, Kyoto University, Osaka University, Okayama University, Kyushu University, Keio University, and Juntendo University, and to store them at the Japan Medical Imaging Database (J-MID) data center. Since the end of 2019 we have been collecting data at full scale and storing it on the RCMB cloud at NII. We have already built up a massive data set, consisting of over 130 million CT images with reports. These data were intended for developing AI tools for diagnostic support, but we wanted to use it to help combat COVID-19 too.

SATOH The RCMB is a research center set up within the NII in 2017 and funded by the Japan Agency for Medical Research and Development (AMED), for the purpose of building a medical big-data cloud network and serving as a hub for AI-based medical image analysis research. A feature of the RCMB is that its research teams are organized collaboratively with other major Japanese research institutes. Researchers in the fields of machine learning, deep learning, networks, cloud computing, and security work together on various kinds of analyses. The collection of real medical images, which are vital for analysis, is currently provided through six medical societies, including the JRS. The images are anonymized and then fed into the NII's medical image big data cloud system for image analysis.

With the continued spread of COVID-19, we have already received CT images of 128 COVID-19 patients from seven facilities connected online to J-MID, and we continue to get new images daily. The Self-Defense Forces Central Hospital, which owns the CT images from the Diamond Princess cruise ship incident, also provided us with scans of 240 cases for AI development.

—— So, all these images become training data for deep learning?

SATOH Yes. To use deep learning for CT image analysis or any other kind of image analysis, you need a lot of training data. For reference (correct diagnosis) data in this effort, we used images of real patients, positive/negative PCR test results, and diagnosis results by radiologists. So far, the AI tool we developed can make a correct diagnosis with 80% accuracy.

As the number of cases increases, we will be able to increase the accuracy further. However, we found that since the PCR test often generates false negative results (when the patient is actually infected), using PCR tests results as reference data does not lead to higher accuracy. For this reason, we are currently using diagnoses made by diagnostic radiologists for reference (correct diagnosis) data.

Overcoming the imbalance of extremely low outlier data

—— Less than 400 cases seems a small number for deep learning. Yet, diagnosis is quite accurate. Is that right?

SATOH For training data, it is also important to have images that are clearly not COVID-19 pneumonia. J-MID has a huge quantity of data from before November 2019, but to make our AI smarter, images with shadows on the lungs that are not cases of COVID-19 are especially valuable as difficult training data. Accuracy can be improved not just by using more images of COVID-19 infections, but also by utilizing tens of thousands of other images as "background" big data.

Inevitably with medical data, there will always be far less case data than normal data. One of the major themes of our research is to improve the accuracy of such unbalanced data by devising various algorithms.

—— Is your AI tool being used already on the medical frontlines?

AOKI We want to see the system widely implemented, but before it can be used clinically we need to go through various procedures, such as pharmaceutical and ethics board approval. We also want to first increase the precision of the tool through verification experiments at universities participating in J-MID. At Juntendo University, for patients suspected of having COVID-19, radiologists look over CT scans closely and for long periods. If they feel the likelihood of infection is high, they proceed with the treatment prescribed for COVID-19 patients, without waiting for a PCR test result. If AI can serve as a useful substitute for a radiologist poring intently over a CT scan image, this would greatly reduce the burden on doctors in the field.

The development of image analysis AI for COVID-19 pneumonia is proceeding rapidly both in Japan and overseas. Some tools have already been approved by pharmaceutical regulators. However, none of these are very accurate. While it is good that regulatory approval was granted quickly in the COVID-19 situation, a lot of verification still needs to be done before such tools can be used confidently in clinical settings.

Using big data for early detection of epidemics

—— Another unique initiative of the Japan Radiological Society relating to COVID-19 is called "viral pneumonia surveillance"

AOKI This is an attempt to get radiologists who observe viral pneumonia on a CT scan to assess the image for signs of COVID-19 and then report the case to us via the JRS website. This is a simple system that does not involve sending images; only the radiologist's findings are reported. We have already collected reports on 1,300 patients. If we plot the results as a bar graph against time, the shape of the graph corresponds closely with a plot of the number of positive PCR cases (issued by the Ministry of Health, Labour and Welfare), despite the fact that the two graphs are based on completely different tests and data collection methods. You can see the second wave fading out now. It's also evident that compared to the first wave, more people in the second wave have relatively mild symptoms.

Another interesting characteristic is that the rise in viral pneumonia patients in the early stage of the first wave occurred ahead of the rise in positive PCR cases. This suggests that this system may be useful for early detection of unknown infections. (See Fig. 2.)

For now, the data are entered manually, but in cooperation

with NII, we want to make it possible to determine and input the likelihood of COVID-19 pneumonia automatically, by simply uploading CT images to the cloud. Also, to get a fuller picture of the spread of infection, we need increase the volume of data. We will do this by cooperating not only with the seven universities connected to J-MID, but also with medical institutions designated for infectious diseases. We hope that with sufficient data, we can prove the system's effectiveness in detecting unknown infectious diseases and in stopping the spread of clusters at an early stage.

The potential of medical imaging big data for healthcare support

—— Aside from COVID-19, RCMB has been working on analyzing various other kinds of medical images.

SATOH Together with the JRS, we are working on the utilization of CT images for detecting subarachnoid hemorrhages and kidney cancer, as well as on the development of AI for estimating and visualizing blood vessels without a contrast media. One unique project we are tackling is a study of brain tumor imaging using magnetic resonance imaging (MRI). Since there are very little data of images of pathological brains compared to normal ones, we are increasing the number of pathological images by using Generative Adversarial Networks (GANs), a deep learning technique used for generating fake images, to create highly realistic images with lesions. This research aims at increasing the accuracy of AI by increasing the number of pathological images. Another challenge we are focused on is acquiring new anatomical knowledge, by using large numbers of CT pelvis images to investigate how the pelvises of Japanese people differ by age and sex, and also to learn more about posterior pelvic tilt, a condition associated with back pain.

Incidentally, the first project we initiated, on detecting subarachnoid hemorrhages, is still under development. This is partly because the appearance of subarachnoid hemorrhages varies enormously, with some images showing no bleeding at all to the untrained eye. It is so difficult to grasp the characteristics of hemorrhages that the AI is not yet sufficiently precise.

AOKI Initially, we medical researchers had excessive expectations for AI (laughs). Subarachnoid hemorrhage is a disease that should never be missed, but if there is no specialist on duty at night, it can end up being overlooked because it is difficult to judge. If an AI tool were able to detect subtle symptoms, the doctor on duty could consult with the AI and proceed to the next treatment without having to contact a specialist for guidance at night. If realized, this would be a very useful application of AI.

In any case, it was great to have so many discussions with the RCMB in the course of joint development. We were able to think together about defining problems to be solved by AI, and also to better understand what AI can and cannot do well. One characteristic of CT and MRI images is that they can capture a wide variety of information all at once, so they are likely to offer more useful information that can be read from them. For example, kidney cancer can be discovered by accident when performing a liver cancer examination. Also, since all the data are digital, they can be fed easily into an AI system. Japanese medical institutions have a massive volume of high-quality data, so if we can use this resource skillfully, we could create AI systems for worldwide use. I hope that the NII will present various use cases to encourage even further data collection and utilization.

SATOH Obviously, we need to show that AI can be useful in clinical settings. We have already developed an AI tool capable of detecting stomach cancer from pathological images. It is already in use at a network of healthcare facilities in Fukushima Prefecture, and we hope to present use cases to follow this. In addition to treating specific diseases, we can utilize medical image big data to develop a system for detecting abnormalities, like the above-mentioned viral pneumonia surveillance initiative, for early detection of epidemics and diseases of unknown infectious pathology. Looking ahead, we want to accumulate more data and accelerate our research on the development of AI to support medical image diagnosis that is more clinically useful and on building better healthcare infrastructure.

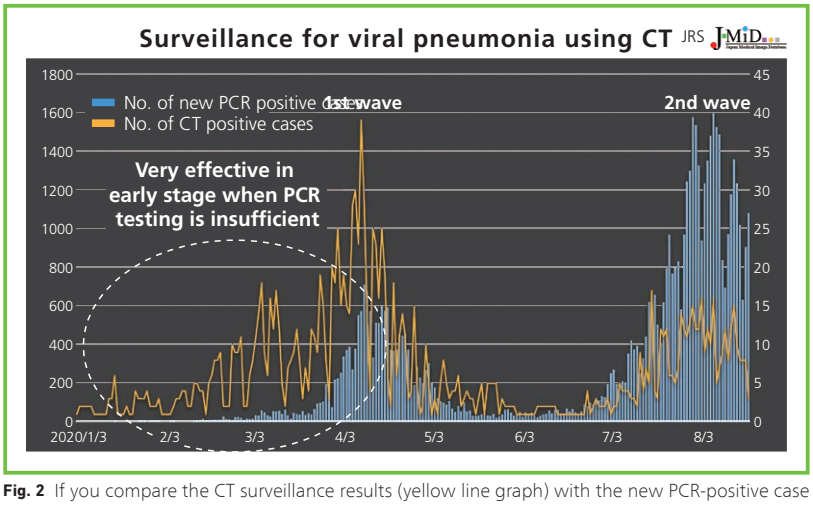


Fig. 2 If you compare the CT surveillance results (yellow line graph) with the new PCR-positive case numbers (blue bar graph), it's clear that their rise and fall correspond closely to both the first wave and the second wave. Although it is unclear whether a limited number of PCR tests can capture the whole picture of an epidemic, CT scans, which are quick and easy to perform, can produce a large quantity of test data that can be comprehensively utilized. This makes CT scans a very useful indicator for capturing the spread of infection in the early stage of an epidemic, when PCR testing may be inadequate.

A Word from the Interviewer

Of the deep learning techniques that sparked the "third AI boom," image analysis is the area that has generated the most dramatic results. Under the right conditions, the AI error rate is now less than that of humans. AI for medical image analysis is expected to be especially valuable in assisting specialists to make diagnoses. However, whether the collection of medical data and the application of AI-based diagnosis continue to advance depend largely on the people at the receiving end. Amid the COVID-19 crisis, many people will have realized that pursuing a "zero risk" approach is not workable. The fact that human lives are at stake makes this a difficult issue, but rather than blindly fearing an AI singularity, humans would do better to improve their scientific literacy and embrace AI as an excellent partner.

TAINAKA Madoka

A senior editor of "NII Today" since 2012. Graduated from the Faculty of Law, Chuo University. Has worked as editor-in-chief of science and technology information magazine "Nature Interface," and as an expert member of the Information Science and Technology Committee of the Ministry of Education, Culture, Sports, Science and Technology's Council for Science and Technology. She is professionally active in the fields of book editing, writing, and PR media for universities and companies. She co-authored a book with KAWARABAYASHI Ken-ichi of NII, titled "This Too Was Mathematics—Car Navigation, Route Maps and Social Media" (Maruzen eBook Library). She pursues the role of interpreter to convey the words of experts in an easy-to-understand way.

Building a Website to Offer a Global Overview of COVID-19 Information

A collaboration between natural language processing researcherspneumonia

AIZAWA Akiko

Professor / Vice Director-General, Digital Content and Media Sciences Research Division, NII
Professor, Department of Computer Science, Graduate School of Information Science and Technology, The University of Tokyo
Professor, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies

Interviewer: HAMAKADO Mamiko

Editor-in-Chief, Natural Sciences and Mathematics (Editorial Dept.), Iwanami Shoten

In reaction to the COVID-19 pandemic, researchers in the field of natural language processing used their expertise to rapidly build a website that aggregates a wide range of COVID-19-related information from around the world and presents it in Japanese. How was this collaboration executed? What was gained by the project? To find out, we spoke to Prof. AIZAWA Akiko of NII, who participated in the project from its inception.

A voluntary initiative of the research community

—— What kind of content does this "COVID-19 World Information Watcher" website offer?

AIZAWA It is easy to understand the structure of the site if you think of it like a table. There are two axes (items), vertical and horizontal. One represents "countries/regions" such as Japan, China, USA, Europe, Africa etc., while the other represents "categories" of information, such as "Current state of infection," "Prevention and regulation," and "Medical info such as symptoms, treatments, and vaccines." (See Fig. 1.) When you visit the site, you see an array of windows organized by "country/region" with a list of linked articles for the selected "category." In fact, there is also a time axis. Information is collected in close to real time, and the articles are listed from newest to oldest in each window.

—— How did this project get started?

AIZAWA It all began with a suggestion by NII Director KITSUREGAWA Masaru. He thought that since we have so many natural language processing researchers, it would be good to do something to help combat the pandemic. After a number of researchers responded positively to the idea and started discussions, a project team took shape.

After the Great East Japan earthquake, we saw young researchers voluntarily launching initiatives to collect information about the disaster, but I think this is the first time in the field of natural language processing that research groups from around the world have worked together so quickly and at such a large scale. At the main international conference in this field, a rapid succession of COVID-19 workshops were held and more than 60 papers were collected, many more than the organizers expected.

Different research labs share the work of building the site

—— How did you design the website?

AIZAWA Professor KUROHASI Sadao of Kyoto University and Prof.

AIZAWA Akiko

Fig. 1 The "COVID-19 World Information Watcher" website: https://lotus.kuee.kyoto-u.ac.jp/NLPforCOVID-19/

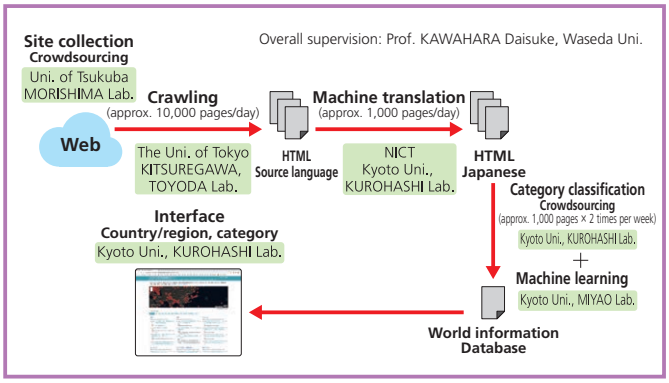


Fig. 2 Flow of website building process and task assignment (see p. 7)

KAWAHARA Daisuke of Waseda University made a rough sketch and then we discussed it all together. (See Fig. 2.) The questions about what to plot on the two axes I mentioned earlier; that is, what countries to collect information from and the categories for organizing information were decided by trial and error.

What was the flow of the website building process?

AIZAWA The first thing we did was to decide the information sources. To collect trustworthy information, we created lists of sites where we could regularly collect up-to-date information, mostly government sources. We also made use of crowdsourcing by choosing sites that were confirmed to be reliable by people in that particular country or region.

The idea was to crawl the selected sites to collect the latest information, but page crawling requires a high degree of skill. Since site content is updated frequently, we needed to check it daily. It was also necessary to deal with the reference content embedded on pages. After collection, the crawled pages, all in different languages, were run through a machine translation engine to convert them into Japanese. As you know, today's machine translation services are very powerful, and certainly good enough to give readers a decent grasp of a document's content.

Once we had the article pages in both the original language and Japanese, we sorted them into "categories" for each "country/region." There were so many articles that we are now using a machine learning process that sorts the articles almost entirely automatically. Finally, the database that we ended up generating in this way is displayed using a web interface, as you can see on the site.

Although none of these elemental technologies were conceived or developed specifically for this project, the final result is more than the sum of its parts. When Watson, an AI computer system developed by IBM, defeated the human champion on a U.S. quiz show in 2011, the achievement lay not in its technological components but rather how they were combined.

Outlook for developing the website further

How do you plan to continue developing the website?

AIZAWA We may develop a variety of modules and incorporate them into the site. One of the nice things about building a site by hand is that you can freely try out functions. There are several possibilities that we have been discussing. It would be nice to have a time series analysis, for example. A constant flow of new articles is steadily added to the site, so it would be great if we could visualize what topics were hot around a particular date. This would give us a "big picture" view of things.

Another possible direction for development is to link the site to a social media platform like Twitter, or to sources of expert knowledge, like scientific papers.

Significance of project and issues that have emerged

What was the main significance of this project?

AIZAWA The project began with the aim of providing people with something useful, but building the site has also been valuable for the project participants. In the process, they have learned about the needs and demands of website users, about what is lacking in current technology, things like that. The value of research lies in acquiring this kind of knowledge. As for COVID-19, the U.S.-based Allen Institute for AI collaborated with various research institutes and organizations on building a dataset of papers on the subject. One of the issues that emerged from text mining this dataset was the lack of any glossary of technical terms relating to COVID-19. To extract technical terminology from papers, we need manually tagged annotated data, but we don't have enough of it. Without the annotated dataset to serve as "scaffolding," detailed analysis cannot be done.

Although the field of life sciences enjoys highly developed technical dictionaries and ontologies, when a pandemic like this suddenly occurs, they are revealed to be insufficient. The technical challenge now is how to create a way to perform efficient and reliable annotation using only a small quantity of data.

Getting back to the "COVID-19 World Information Watcher" site, we had a surprise related to annotation there too. At first, I thought we could do almost anything by crowdsourcing, but it turned out, naturally enough, that we did not have enough "crowd workers" for every language, region, and category. Given that the success or failure of an AI system in a particular domain depends on securing enough people who can annotate data for that domain, this problem can lead to very large biases and disparities. This is something I learned only after dealing with massive quantities of data arriving simultaneously from real human communities around the world.

A Word from the Interviewer

It seems to me that this kind of website, which offers a "big picture" view of information from across the whole world in Japanese, should be used above all by political and administrative authorities and the mass media, as a resource for careful policymaking and reporting.
In the text, there is a reference to increasing the accuracy of machine translation. In preparing this manuscript, I made use of an automatic speech-to-text conversion service, finding that this technology has now entered the realm of practical usability. The benefits of such improvements in machine translation, as well as natural language processing technology, are more and more directly evident.

HAMAKADO Mamiko
Editor-in-Chief, Natural Sciences and Mathematics (Editorial Dept.), Iwanami Shoten
After graduating from Tohoku University with a B.Sc. and from The University of Tokyo with an M.Sc., she joined Iwanami Shoten in 1993, where she was assigned to work in the Natural Science and Mathematics editorial department, taking charge of publishing books on cognitive science, linguistics, and mathematics. She has worked in her current position since 2015.

Commentary

Research Through Open Collaboration in the Post-Covid Era

KAWAHARA Daisuke

Professor, Department of Computer Science and Communications Engineering, School of Fundamental Science and Engineering, Waseda University

Received a B.Eng and M.Eng from Kyoto University in 1997 and 1999, respectively. Withdrawal from Kyoto University doctoral program with the completion of course requirements in 2002. After working as a research assistant at the University of Tokyo, a senior researcher at the National Institute of Information and Communications Technology, and an associate professor at the Graduate School of Informatics, Kyoto University, he assumed his current position in 2020. Engages in research on natural language processing and knowledge processing. PhD (Informatics).

In early March 2020, COVID-19 started to spread in Japan, after causing havoc in China, South Korea, and Italy. In reaction to this situation, NII Director Prof. KITSUREGAWA Masaru asked the natural language processing researchers working on the JST (Japan Science and Technology Agency) "CREST Sakigake" multi-disciplinary "Big Data Infrastructure" research project (led by Prof. KITSUREGAWA) if they could do something to help deal with COVID-19. The situation could be described as a "national crisis," or even a "global crisis," so naturally we all wanted to apply natural language processing technology for the benefit of society. That is how this project (see pages 5-6) was launched by a group of volunteer researchers.

Unfortunately, the COVID-19 crisis did not allow for face-to-face meetings. From the beginning of the project, therefore, we made use of Zoom for online meetings, Slack for daily communication, and GitHub for code sharing.

Looking back over the Slack logs from those early days, I noted that the ideas we raised were as follows.

- **Fact checking**
- **Literature analysis: Knowledge discovery from COVID-19-related articles**
- **Collection and translation of information in the various languages of different countries**
- **Verbalization of numerical data such as infection numbers and change point detection**
- **Time series event analysis**
- **Recognizing problems and where they were occurring**

As we discussed these ideas, we realized that other research groups were already working on fact-checking and article analysis, so we decided to think of something else that might be useful. That is what led us to collect information (web pages) from countries all over the world, convert it into Japanese using machine translation, and classify it into categories. We also settled on the idea of structuring this information in the form of a website, called "COVID-19 World Information Watcher," with a basic two-dimensional presentation by "country" and "category."

A big focus of the discussions at the time was the target demographic of the website. Bureaucrats, politicians, corporate executives, medical professionals, and the general public were among the suggestions, but a conclusion was not reached. In the end, we decided to start building the site and to figure this out along the way.

By late March we had completed a prototype. For machine translation, we initially used Google Translate, but the cost of translating large numbers of pages became excessive, so we switched to the free "Minna no Honyaku" (NICT). After some negotiations, we were granted unlimited access, which was very helpful.

To classify the information, we set up six categories, such as "current state of infection" and "economic and welfare policies." Initially, we also created a classifier based on keywords for collecting data. However, since categorization and translation were automatic processes, the accuracy was limited. As always with this type of technology, if a service does not meet a minimum level of accuracy, people will not use it. By May, we started crowdsourcing for category annotation. We asked several "crowd workers" to assign categories to each page. By aggregating these results, we attained a high level of accuracy. This arrangement also ended up being too costly, though, so after two months we terminated it.

Later, we managed to improve classification accuracy with machine learning, using the generalized contextual language model BERT and feeding in the previous category assignment data as training data. We also developed a system for manually correcting incorrect categories, sharing the work between ourselves. In the end, human resources played a crucial role.

After all this hard work, we launched the website to the public in June. The rapid construction of the site was made possible by the fact that all the researchers took full advantage of their respective technical specializations and divided the tasks between them. I played the role of a director, integrating modules arriving from the various labs.

Further challenges include delivering the website to users who most need it and improving the interface design to make the process of finding useful information more efficient.

BERT is the tool we used for classification. BERT training consists of two stages, pre-training and fine-tuning. Pre-training is computationally expensive, but anyone can do this once and publish their results. Resources like this are rapidly being released and shared in English and Chinese-speaking countries, while Japan lags far behind. It is therefore necessary for researchers and labs to divide tasks and collaborate on a national scale. Challenges like this one are vital. I believe that this kind of open collaboration across organizational boundaries is not only very well suited to the post-COVID-19 era, but also increasingly important.

Exploring Social Media's "Fake News" Problem

The importance of questioning information sources

TORIUMI Fujio

Associate Professor, Department of Systems Innovation, Graduate School of Engineering, The University of Tokyo

Interviewer: MURAYAMA Keiichi

Commentator, Nikkei, Inc.

With the continuing spread of COVID-19 infections, the propagation of "fake news" through social media has become a concern throughout the world (See Fig. 1.). In Japan, for example, false rumors leading to a toilet paper shortage drew widespread public attention. How should we view and understand this phenomenon? We asked TORIUMI Fujio, an associate professor at The University of Tokyo, about the issue of "social media and fake news."

People give more importance to what's funny and interesting than what's correct

—— What lies behind the phenomenon of "fake news" and how does it start?

TORIUMI There are several reasons. In most cases, the original intention is not to deceive people. More often there is some misunderstanding behind it. As soon as something happens, people tend to look for a cause-and-effect relationship. As a result, we get misunderstandings and the propagation of rumors. The reason why such information gets circulated is not that people are in a panic. Most people do not even think about whether a particular piece of information is correct or not. Information generally spreads without any particular suspicion attached to it. That itself is a very natural process, I think.

Sometimes, people spread information because they find it interesting. Even if the content is questionable, they share it because it gives others something to talk about, or it stimulates conversation between friends. It's more interesting to post "I heard that toilet paper is disappearing" than "I saw toilet paper on the shelves today." So, rumors and fake news spread because people are more inclined to share amusing thoughts than to convey accurate information. This is the second reason for the spread of fake news.

We should also note that the mainstream media sometimes spreads false information too. There are several patterns of news coverage that lead to this, but a typical example in the context of COVID-19 situation is reports that refer to certain stories going viral on social media, when in fact the stories are far from viral. There have been cases where the mass media reported that such-and-such fake news was being spread on Twitter, when only a handful of people actually wrote anything about the topic.

Received a PhD (Engineering) from the Department of Control Engineering, Graduate School of Engineering, Tokyo Institute of Technology in 2004 and has worked in his current position since 2012. He pursues research on computational social science and social applications of AI technology. He was recognized for his significant contribution to science and technology in 2018 by the National Institute of Science and Technology Policy ("Nice Step Researcher").

TORIUMI Fujio

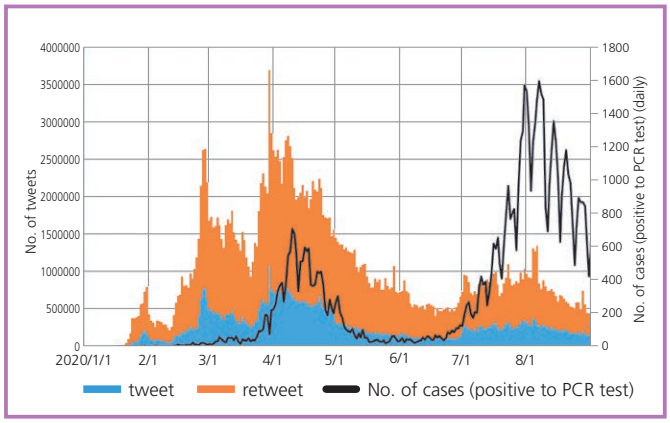


Fig. 1 Change over time in tweets about COVID-19 during the pandemic

—— How do you see the impact of fake news related COVID-19?

TORIUMI Apart from the toilet paper shortage, which was the hottest topic, there were reports claiming that drinking hot water and taking Isodine were effective against the virus. There has certainly been fake news, but I wonder how much it really spread. Quite honestly, I don't think there were many instances of fake news being deliberately propagated. And in any case, I don't think it spread very much.

Admittedly, it is difficult to determine how much amount of fake news is considered "spread." Take for example the toilet paper rumors. A detailed analysis showed that the number of people who might have seen tweets to the effect that toilet paper was running out or might run out was very small (approx. 2%)

compared to those who saw tweets cautioning that the news was fake. We can safely conclude that only a very limited number of people would have only seen the fake news. One person was accused of starting the false rumors, but evidence indicated that what the person wrote did not really spread. The post in question was retweeted once or twice. And since the tweet was self-initiated, there was effectively zero spread. (See Fig. 2.)

Why did a toilet paper shortage occur, then? You might reason that it was because people fooled by the false rumor rushed out to buy some, fearing stocks would run out. However, the people buying toilet paper were not young people, aged in their teens to 30s—the main users of Twitter—but rather the older generations that barely use Twitter at all. It is therefore quite doubtful that this phenomenon was caused by fake news originating on social media.

The problem is not the "amount" of spread

—— Do you think we should reject the notion that the spread of fake news on social media has caused social confusion?

TORIUMI Overwhelmingly, it is the young generation that uses social media. A survey on COVID-19 information sources showed that the proportion of people who get their information from social media is high in the age range of teens to 30s, about 25% for people in their 40s, and less than 10% for people in their 60s. Since the age distribution of the Japanese population is skewed to the older side, it would be hard to argue that the influence of social media is strong. Another point is the question of trust. Even in the young age group, the level of trust in social media is less than 20%. Few people including those in the young age group see Twitter as a trustworthy source of information.

At least in Japan, we don't have situations where fake news spreads to the point that everyone is fooled. At the same time, it's true that false rumors can be dangerous even if they don't spread on a large scale. There is a certain amount of denial of scientific ideas, which can be life-threatening in some cases, and this problem cannot be quantitatively measured. We cannot definitely say how much or how little certain news spread. But even a small-scale outbreak of false rumors can result in fatal consequences. As in the case of COVID-19, fake news is more likely to be a problem in times of disaster or emergency.

Is spreading information a social contribution in times of emergency?

—— Is there a difference in how much fake news is spread during emergencies compared to normal times?

TORIUMI COVID-19 and natural disasters are topics that everyone gets interested in at the same time. Basically, people on social media form their own communities to access the kind of information that interests them. They are active in their own areas and often use social media for fun. Since different communities usually have little in common with each other, not much information spreads across communities. However, with something like COVID-19, everyone is simultaneously interested in the same thing, creating fertile ground for the diffusion of fake news.

During an emergency, people also consider it valuable to spread information that seems important. This is especially true for disasters, when people see sharing information as a social duty. They get a sense of satisfaction because they feel they have done something good for society. In other words, with little or no effort they can feel that they have contributed to the community.

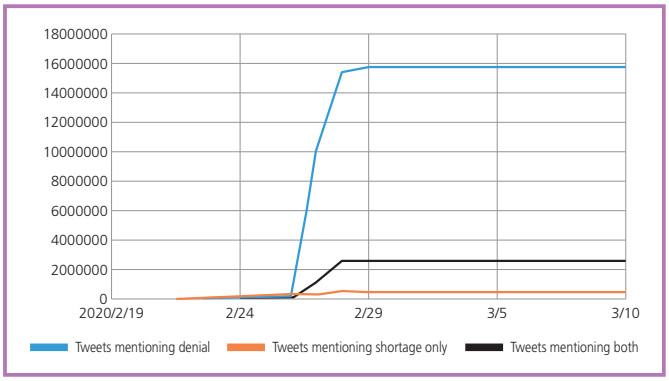


Fig. 2 Spread of tweets denying and reporting toilet paper shortage

Since this kind of behavior becomes common in a disaster, it is also easier for false rumors or spurious information to spread widely.

—— In light of this situation, what should we do about social media use?

TORIUMI It is dangerous to assume that just because there have not been any problems in the past, it is fine to go on as before. Then again, it is difficult to question the validity of all information. Personally, I think the world tends to run more smoothly if we take a positive view of human nature. Nonetheless, there are several points we should be skeptical about. For example, it's important to question sources of information. "Is this person suspicious?" "Are their retweets biased?" In many cases, fake news is communicated for selfish motives or personal benefit. So, it is always vital to check the source.

To put it another way, it is important to have meta-information about the source of information and who is spreading it. Another criterion for making judgments is the trustworthiness of media sources. If we can just pause to consider whether there is anything suspicious about information, when we encounter it, we can defend ourselves against it. In this sense, providing meta-information to help this kind of questioning would be helpful. With some background information, people can investigate the veracity of information for themselves and learn that alternative viewpoints may exist.

A Word from the Interviewer

When I spoke with Jack Dorsey, CEO and founder of Twitter, he explained that Twitter is being used beyond the company's initial expectations. His overall assessment was positive, saying that he felt "touched" to see people using Twitter for information gathering and communication. At the same time, he added that we cannot neglect to deal with the issues of misinformation, disinformation, and hate speech. We are not yet bringing out the full potential of social media. There is plenty of room for more intelligent use.

MURAYAMA Keiichi

Joined Nikkei, Inc. in 1992. His work in the Industrial Division focuses on information technology and electronics, automobiles, healthcare, and finance.
He studied at Harvard University in the USA from 2004 to 2005 and served as Nikkei's Silicon Valley Bureau Chief from 2005 to 2009. He became an editorial board member in 2012, and an editorialist in 2015. Since 2017 he has worked as a commentator, covering IT and startups.

Big Data Analysis for Resolving the Social Problems of COVID-19

Visualizing the "self-restraint" rate for going out and human contact frequency

MIZUNO Takayuki
Associate Professor, Information and Society Research Division, NII
Associate Professor, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies

Interviewer: YAMAMOTO Kayoko
Editorialist and Editor, Nikkan Kogyo Shimbun

MIZUNO Takayuki, an associate professor at NII, makes use of his knowledge of big data analysis and large-scale simulations, cultivated in the fields of informatics and physics, to try and solve economic and social problems. As a data scientist specializing in "computational social science," he has been analyzing big data from mobile phone networks, GPS, and other sources, and actively sharing information to help combat COVID-19. On top of this, he is striving to use AI (artificial intelligence) to capture the constantly changing infection conditions and circumstances, which vary from country to country.

Making use of data analysis results for policymaking

—— In the past few years, there has been a lot of interest in research on visualizing flows of people and business activities in real time by means of "big data" analysis of location, purchase history, etc. from mobile phones. There are hopes that these data can also be used to help combat COVID-19.

MIZUNO On one side, there are vendors that collect data, such as mobile phone networks and other private companies. On the other side, there are national and local governments that make policy decisions based on the results of data analysis. In between the two are data scientists, who process essential parts of big data sets to extract meaningful and useful information. Since there were no data scientists working on the COVID-19 issue when data started being used to develop countermeasures, there was initially some confusion.

Data analysis itself can be handled by data experts from other fields, but if it is not done with the right methods, there is a risk that arbitrary or biased results will get published. Data also have their own quirks, especially when used for official statistics. Take mobile networks for example. NTT DoCoMo has an older user base than Rakuten Mobile, and the proportions of these users vary widely from one area to another. To do justice to data analysis, a data scientist who is familiar with data and analysis techniques is indispensable. Researchers with expertise in informatics including those at NII needed to take the lead in getting involved. So, in mid-March, we began researching COVID-19 countermeasures, in collaboration with infectious disease researchers.

MIZUNO Takayuki

"Self-restraint" from going out is about "outflow" not "inflow"

—— When the government asked people to practice "self-restraint" (jishuku) by not going out as part of its first COVID-19 countermeasures, with the aim of reducing human contacts by 80%, the number of people in downtown streets dropped. At the same time, though, the number of people in neighboring shopping districts and parks increased.

MIZUNO This happened because we were focusing on the wrong data set. To get people to refrain from going out, the vital point was to control the "outflow" of people from their residential areas; not the "inflow" of people to commercial areas.

So, a group of us from NII and The Canon Institute for Global Studies conducted a study along these lines. Using real-time population distribution information estimated from base station data for around 80 million DoCoMo mobile phones, we estimated the number of people going out from residential areas (nighttime minus daytime population). We then derived the "self-restraint" rate using the ratio of the number of people going out during the pandemic to the number going out in normal times. More specifically, we visualized the degree to which residents in specific municipalities refrained from going out (relative to the pre-COVID-19 period) using the formula "1 – (no. of people out on a certain day)/(no. of people out in normal times)." (See Fig. 1.) When we published this statistic on the NII website, NHK (Japan Broadcasting Corporation) reported on it for several months, as a way of encouraging good behavior. We also verified the correlation between "self-restraint" rate and infection suppression, as well as differences in behavior by age and gender.

Although we didn't make it public, we also estimated human outflows for the cells of a 500-meter square mesh. This enabled us to correlate the "self-restraint" rate figures with local features, such as the presence of childcare facilities, factories, and shopping

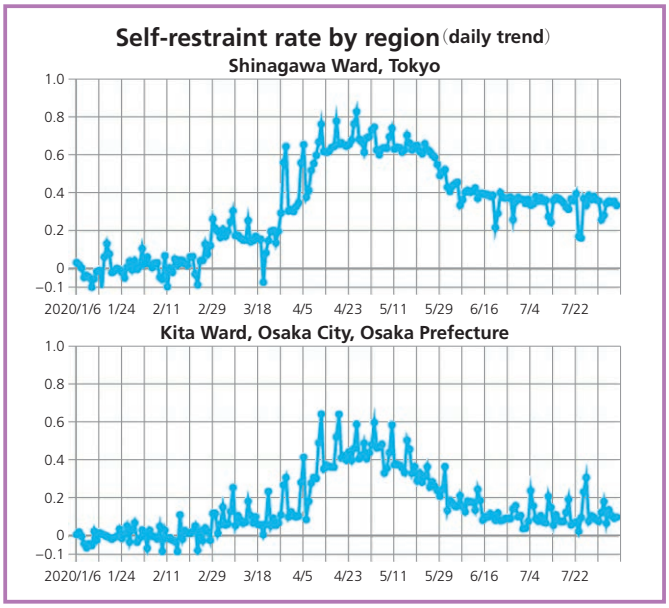


Fig. 1 Visualization of "self-restraint" rates of local residents using big data on flowing populations. Self-restraint rate is searchable by prefecture and region.

areas. Prefectural governments, and city and ward governments in the case of Tokyo, can now utilize these data to back up their requests for "self-restraint."

In Tokyo at the peak of the crisis, the "self-restraint" rate was as high as 60-80%, and as of mid-August the outflow from residential areas in Tokyo was still 20% below normal. This is likely due to telecommuting. The results of this analysis are now being used to investigate correlations with economic activity.

What needs to be controlled for reducing contact frequency

—— Later, the government's policy changed from a blanket request for self-restraint to the practice of social distancing for all social activities.

MIZUNO That's why we focused on contact frequency, which leads to person-to-person transmission. We used big data on GPS locations to generate high-resolution population distributions. To do this, we combined two types of data—the location data of about 80 million people from DoCoMo base stations at a resolution of 500-meter square cells and location data collected from some 200,000 consenting mobile phone users, accurate to several meters. That is, we combined a huge quantity of low-precision data points with a smaller quantity of high-precision data points.

We need to note that GPS only provides 2D information. In the case of tall buildings with people on each floor, the human density is not actually high. And even when density is high, if most people are traveling by car, there is no problem. We even considered things like the direction of people's movement. The real problem is when people are close and face-to-face with each other for longer than 15 minutes. Taking into account these factors, we estimated the change in population density per unit of time and the change in the number of people contacted per person. From this, we derived a fairly realistic estimate of contact frequency. (See Fig. 2.) The analysis results proved that to cut human contact by 80%, it was not necessary to cut the number of people going out to the city by 80%; a reduction of just 65% was sufficient to achieve this goal.

Using AI to find the X factor

—— There is now a need to establish a model to predict infectious disease outbreaks. There are various factors involved, like the "self-restraint" rate and the distribution of people. Can such a model tell us how to effectively control the spread of disease?

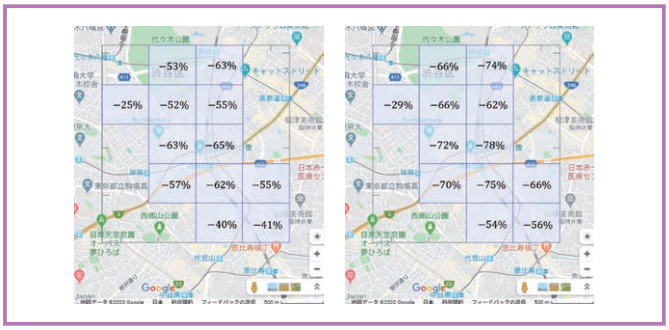


Fig. 2 Daytime data for a weekday (April 24) in the Shibuya district (compared to January 17). The population change rate (left) for 500-meter square cells and contact frequency change rate (right = population density effect q=0.68) are calculated. The contact frequency provides a more realistic picture of density conditions. This kind of analysis makes it possible to tailor countermeasures for specific regions.

MIZUNO There are two basic techniques to derive such models. One is to model the epidemic based on a classical physical model, using only important factors with well-known causal relationships based on economic and epidemiological theories. In this case, it is easy to understand which factors contribute to infection, how and to what degree. Due to the small number of factors, however, accuracy is not high. On the other hand, with AI based on machine learning, we can use many factors to predict infection with much greater accuracy. In this case, though, the model is a black box, so we cannot easily understand which factors are effective and how they work. Still, this approach can be useful for discovering unexpected and unknown factors.

As an example, consider the surprising incident of healthy elderly people enjoying karaoke at a daytime bar, resulting in an outbreak of infection. AI would probably be able to detect such locally specific factors that are difficult to notice. I would like to use AI to investigate the reason behind the difference in infection rates between Asia (including Japan) and Europe and the USA. What is the "X factor" behind this difference, I wonder. This kind of analysis is also not limited to COVID-19. It could be useful in understanding all kinds of phenomena occurring around the world.

(Photography by SATO Yusuke)

A Word from the Interviewer

Researchers who conduct big data analysis and surveys generally avoid sending strong messages when they publish their results. This is an understandable result of caution, when all the influencing factors have not yet been studied, but it is also due to the reluctance of researchers to speak about areas outside their field of expertise. In response to this, Prof. MIZUNO explains, "The model we used has its limitations, and there will always be errors." I look forward to following the activities of Prof. MIZUNO and NII, as they actively tackle complex social issues with scientific integrity.

YAMAMOTO Kayoko

Editorialist and Editor, Nikkan Kogyo Shimbun. After earning a master's degree at the Tokyo Institute of Technology in 1990, she joined Nikkan Kogyo Shimbun, working on science and technology and the chemical industry. She has now specialized in reporting on "university-industry collaboration" for almost 20 years, and earned a PhD from the Tokyo University of Agriculture and Technology for research in this same field. She is member of the Kisha (Press) Club of the Ministry of Education, Culture, Sports, Science and Technology.

Essay

A Time of Change in Education

YASUURA Hiroto
Trustee and Executive Vice President, Kyushu University

The suddenness of COVID-19 took the world by surprise. After beginning in Wuhan, China at the end of last year, by the latter half of February 2020, COVID-19 had spread to Japan, Europe and the USA. By March, it had become a global pandemic. As of the end of August 2020, it had infected 25 million people and led to over 850,000 deaths worldwide.

It is spring. Flowers are blooming, mountains are alive with fresh growth, and seas and rivers are transforming into their summer colors. As familiar streets and landscapes mark the changing season, people's lives have changed dramatically. Festivals and events are cancelled, commuting to school and work is reduced, and downtown city areas no longer bustle. At universities there are no graduation or entrance ceremonies, and all classes are offered online only. Even as the physical structure of society is remains intact, we are suddenly forced to operate in a cyberspace-centered world.

While many people may have imagined that such an emergency might arrive with war or a natural disaster, few would have expected such drastic social change to arrive like this. Yet, if we look at things from a different angle, we realize that the "cyber physical systems" (CPS) that we have constructed are functioning well, enabling a smooth continuation of the activities of society. In particular, I believe that future historians might appreciate the fact that so many educational institutions in Japan and abroad have been able to implement distance learning to keep alive the education and training of young people, almost without missing a beat.

Four years ago, I wrote an essay for this same space, titled "Information Technology is a Bastion of University Reform." Some of the things I imagined in that essay have now come to fruition. In research and university administrations too, there is a growing momentum toward digital transformation (DX).

Especially in the area of DX for education, countries all over the world find themselves under similar circumstances at much the same time. They are compelled to generate various kinds of new ideas, build tools, and establish operational standards and know-how. In Japan too, the NII and major universities have led a series of online conferences titled "Cyber Symposium for Sharing Information on Distance Learning Efforts at Universities for the New Academic Year" a total of 15 times (as of September 4), for the purpose of sharing methods and practical examples of online education from universities to elementary and secondary education level, as well as to present details of the situation overseas. It is likely that in 10 years' time, we will look back at 2020 as a critical moment of fundamental change in the nature of education. This is a global trend that is likely to become a massive shift, encompassing everything from elementary and secondary education to higher and adult education. Standing at this historical turning point, I feel fortunate to have played a small part in the reform of information technology and the educational infrastructure that utilizes it.

References

YASUURA Hiroto: "Information Technology is a Bastion of University Reform",
Essay, NII Today No.72,
https://www.nii.ac.jp/today/72/7.html

Future Schedule

October 17-18 | Inter-University Research Institute Corporation Symposium 2020 (exhibition) = online. See https://ius.4kikou.org/ for registration and details.
Late October | 2nd SPARC Japan Seminar 2020 = online. See https://www.nii.ac.jp/sparc/event/ for details.
November 1-30 | 22nd Library Fair & Forum (exhibition) = online. With core time. See **https://www.nii.ac.jp/event/other/libraryfair/** for details.
December 9-11 | Academic eXchange for Information Environment and Strategy (AXIES) 2020 Annual Conference (exhibition) = online. **See https://axies.jp/conf/conf2020/** for details.

Information

The NII is now accepting applications for public joint research projects for FY2021.

The following three types of joint research are open to the public. The application period is October 1 to December 1, 2020. We are looking for research proposals to help overcome the COVID-19 crisis. Young researchers, female researchers, and researchers from remote universities are also welcome to apply.
(1) Open Call for Strategic Research, (2) Open Call for Research Planning Meetings, (3) Open Call for Free Proposals
See the website for details about application and eligibility requirements. https://www.nii.ac.jp/research/collaboration/koubo/

Notes on cover illustration

Ground glass opacity on a CT image of lungs is characteristic of COVID-19 pneumonia. A metal robot cannot actually undergo a CT scan, but a human can help to confirm the presence of COVID-19 and the condition of lungs. Such CT images are being used as training data for the development of AI-based diagnostic support tools.

Weaving Information into Knowledge
NII

National Institute of Informatics News [NII Today]
No. 75 Sep. 2020
[[This English language edition NII Today corresponds to No. 89 of the Japanese edition.]]
Published by National Institute of Informatics, Research Organization of Information and Systems
Address | National Center of Sciences 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430
Publisher | KITSUREGAWA Masaru Editorial Supervisor | SATOH Ichiro
Cover illustration | SHIROTANI Toshiya Copy Editor | TAINAKA Madoka
Production | MATZDA OFFICE CO. LTD., Sci-Tech Communications Inc.
Contact | Publicity Team, Planning Division, General Affairs Department
TEL | +81-3-4212-2028 FAX | +81-3-4212-2150 E-mail | kouhou@nii.ac.jp

Bit (NII Character)

https://www.nii.ac.jp/en/about/publications/today/