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

NII Today

71
Sep. 2019
National Institute of Informatics News

NII Interview

Teaming Up With Researchers From Around the World to Combat Increasing Amounts of Fake Information

Isao Echizen [NII Acting Director General/Deputy Director General / Professor, Information and Society Research Division, NII]
Junichi Yamagishi [Professor, Digital Content and Media Sciences Research Division, NII]
Xin Wang [Project Assistant Professor, Digital Content and Media Sciences Research Division, NII]
Fuming Fang [Project Researcher, Information and Society Research Division, NII]

Trying Evaluating the Authenticity of Internet Discourse

Kentaro Inui [Professor, Graduate School of Information Sciences, Tohoku University]

Three-Way Talk

Reflection on Research Misconduct Leads to Development of Original Science-Misconduct Prevention System

Yukihide Tomari [Deputy Director / Professor, Institute for Quantitative Biosciences, The University of Tokyo]
Takashi Sutani [Lecturer, Institute for Quantitative Biosciences, The University of Tokyo]
Yusuke Komiyama [Assistant Professor, Digital Content and Media Sciences Research Division, NII]

Feature

Tackling Fakes

To Detect Fraudulent Information

This English language edition
NII Today corresponds to
No. 85 of the Japanese edition

Teaming Up with Researchers from Around the World to Combat Increasing Amounts of Fake Information

Anticipatory R&D in preparation for attacks

Isao Echizen
NII Acting Director General/ Deputy Director General / Professor, Information and Society Research Division, NII / Professor, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies

Junichi Yamagishi
Professor, Digital Content and Media Sciences Research Division, NII / Professor, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies

Xin Wang
Project Assistant Professor, Digital Content and Media Sciences Research Division, NII

Fuming Fang
Project Researcher, Information and Society Research Division, NII

Interviewer: Nobuyuki Yajima Senior Researcher, Nikkei BP Intelligence Group, Nikkei Business Publications, Inc.

Collectively known as "fakes," maliciously manipulated images, audio, and posts continue to increase on the Internet. Fakes can lead to bad decisions, invalidation of information security, and privacy violations, and they are therefore a threat to society. NII's Professor Isao Echizen, Professor Junichi Yamagishi, Project Assistant Professor Xin Wang, and Project Researcher Fuming Fang are discovering protective measures by predicting attack methods ahead of the attackers and using technologies such as machine learning. They are gathering the necessary data to tackle the problem in collaboration with researchers worldwide.

Team research that crosses disciplinary boundaries

—— Why are the Echizen Lab and the Yamagishi Lab working together to tackle this problem of fake information?

Echizen The scope of fake information is wide and includes images, audio, and posted text. Moreover, fake information that combines these is also starting to appear. It is partly because the problem is becoming increasingly so complicated that I, specializing in images, and Professor Yamagishi, specializing in audio, are working together to come up with countermeasures. Originally, I was doing research on protecting contents shown on displays and screens, such as technology for preventing surreptitious filming, but I shifted recently to research aimed at protecting human beings, and I encountered the problem of fake information in that research.

Yamagishi I was primarily working on "smart speech synthesis," creating speech that fits a situation, but as speech synthesis technology has evolved, it has become impossible to avoid the problem of fakes.

My focus is on an international competition aimed at distinguishing between a human voice and a computer-synthesized voice, and I am involved in planning and implementing this competition by calling on researchers worldwide. I am approaching the research by going beyond disciplinary boundaries, including getting participants from the Echizen Lab. This is because, no matter what the subject, the development of machine learning means

that all problem-solving methods are similar. Unless, we view problems with a wider perspective and consider the whole picture, rather than researching speech only or images only, we won't be able to do new research.

Wang I am working on speech research at the Yamagishi Lab, and I am administrating the competition together with researchers from countries such as France, Ireland, and Finland. In addition to the competition, I am engaged in a variety of joint research with researchers from European countries.

Fang I am affiliated with the Echizen Lab, but my research isn't just about images. I'm also researching technology for simultaneously transforming the face and the voice in a video of someone talking, as well as measures to prevent its misuse, and technology to protect privacy by anonymizing the voices of individuals. Speech research requires high-quality speech synthesis, and we are using a method devised by Xin Wang for that.

Devising measures against ever-increasing threats

—— Not only is there collaboration between the two labs, but you are also teaming up with researchers from around the world to tackle the problem of fake information. Is that an indication of how important the issue is?

Echizen There is an expression, "seeing is believing," which expresses the tendency we have to believe things once we have confirmed them with our own eyes, rather than just relying on hearsay. However, nowadays, the information we see could be fake. In other words, "seeing" has become questionable. There are fake images and fake voices that are perfect likenesses, and they can cause people to draw incorrect conclusions. There is also a danger that attackers will break into systems by getting past biometric authentication using fake faces, fingerprints, or speech (Figure 1).

Over a billion smartphones are sold every year, and most people are carrying around high-resolution cameras that they use to take photos. This means that high-quality images can be published on the Internet immediately. This situation lies behind the problem of fake information. Also, it is now possible to take a photo of a stranger and identify him or her using a smartphone facial recognition app, which brings up the issue of privacy violation.

Yamagishi On the topic of speech synthesis, in 2017, speech created by a subsidiary of Google that developed AlphaGo (DeepMind) passed the Turing test. This means that the technology has reached a level at which people cannot tell that a voice has been created by a computer. Software created by Microsoft has also passed the same test. Imagine what risks might arise if someone with malicious intent were to be able to synthesize a voice that sounded exactly like yours without your permission.

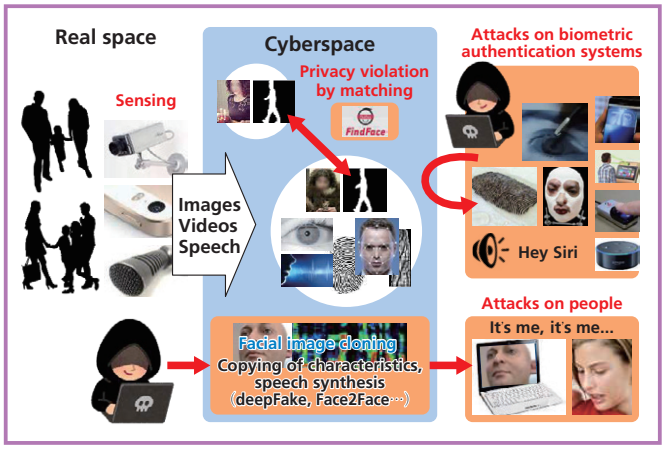


Figure 1 Threat associated with sensing of biometric information
Through many high-precision sensors, our biometric information is shared in cyberspace, and this allows privacy violations by matching, attacks on biometric authentication systems, and attacks on people.

Echizen I think that researchers have a duty to anticipate attack methods and highlight the threats. Of course, they must also prepare specific protective measures.

Yamagishi As the technology advances, it will become possible to do things like enhancing entertainment using image generation, and using speech synthesis to restore the voice of someone who has lost their own voice due to illness in such a way that it sounds exactly like it did before. That is why, rather than saying we won't develop the technology, because there are risks, we take the stance of providing measures to protect against misuse as we evolve the technology.

Specific Measure ❶: The world's first automatic detection of video authenticity

—— What kind of specific protective measures are there?

Echizen We have developed the world's first neural network capable of automatically detecting the authenticity of videos using machine learning. When a video is loaded into the network, it determines whether the video is real or fake.

We collected real and fake videos and taught the network by indicating which were real and which were fake. The network divides a loaded video into several frames, extracts features from

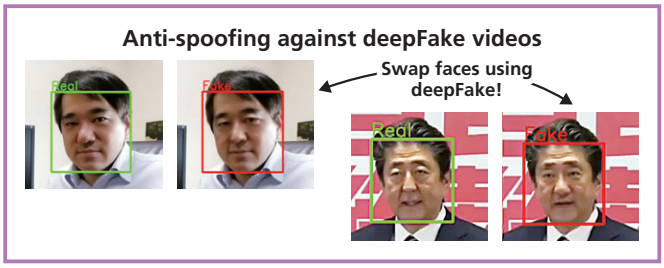


Figure 2 Detection of fake facial images
Example using a technique that automatically determines the authenticity of facial images that have been swapped using deepFake.

each frame, and then combines the results and indicates the presence or absence of manipulation.

In the case of uncompressed video data, we reached a point where we were able to limit the rate of mistaken detection to approximately 0.67–0.77%. Continuing our research, we made it possible for the network to detect information such as which of multiple video manipulation methods has been used to change which parts of the video, for example, the face or the mouth (Figure 2).

Yamagishi If companies engaged in video sharing services were to use this software and place a sign saying "This video has been manipulated" below the videos they release, it might reduce the risk of people believing incorrect information.

Specific Measure ❷: Hosting an international competition for distinguishing between genuine and fake speech

—— What is the international competition for distinguishing between genuine and fake speech?

Yamagishi During the course of our research on speech synthesis and analysis, we have been tackling automatic identification of genuine and fake speech since 2010. Currently, we are developing a system for detecting whether a speaker is human, in other words, a living body. This field of research has gained momentum worldwide, and we have held a competition called the ASVspoof Challenge[1] biennially since 2015. In 2019, 154 companies and research institutions took part in the challenge, and 49 teams submitted detection results. Companies such as Google, NTT, and HOYA, and research institutions from countries such as Finland, Ireland, and France participated.

One feature of the challenge is that it takes an open data approach. Using speech data that I had published in the past, cooperating companies manipulated and synthesized it to provide a variety of synthesized speech data. The challenge participants were given a total of 19 types of natural and synthesized speech data. They analyzed the data using their own technology and submitted the results of their authenticity assessment. The providers of the speech data did not participate in the analysis challenge.

A problem that bedevils machine learning is the collection of high-quality data. Creating measures against fakes requires fake data in addition to correct data. Thanks to cooperation from companies, we were able to give the participants worthwhile data.

The results of the challenge exceeded my expectations. An American information security company ranked first, with an error rate of just 0.2%. This means that the computer can determine authenticity even when the human ear cannot. What was distinctive about this entry was that they prepared multiple detection models and combined them to distinguish between the genuine and fake data. Fakes are becoming more sophisticated, and therefore we need to respond using multiple models by preparing for various methods.

Specific Measure ❸: Preventing "spoofing attacks" by anonymization

—— What about measures to prevent the misuse of biometric information such as facial images and speech?

Echizen We have carried out various research projects on images. For example, when a person wears a special pair of glasses that we developed called PrivacyVisor, their privacy is protected even if their face is photographed because the glasses prevent facial recognition technology from working properly. Also, we developed a measure to prevent fingerprint data being taken from photos in which people are making the peace sign. This method involves transferring a special image pattern that prevents surreptitious photography onto the fingers simply by pressing them onto a sheet, and it has no adverse effect on the fingerprint authentication that people habitually use.

Fang We also developed a speaker anonymization method. When publishing speech data, this method processes the data so that individuals cannot be identified. Even though it is processed, it still sounds like natural speech to the human ear. Changing the voice to a different person's voice or bleeping it out would sound unnatural.

Specifically, we analyze the speech data that we want to anonymize to discover the characteristics, and then we look for another person's speech data that has similar characteristics and synthesize the speech to take the average. Speech data created in this way do not sound strange to the listener, and the individual cannot be identified because the speech characteristics have been averaged.

Specific Measure ❹: Innovation in pivotal speech synthesis "dramatically reduces computational complexity"

Yamagishi Whether it's anonymization or creating data for the analysis challenge, the key is whether high-quality speech synthesis is possible. Synthesis technology is fundamental, not only in measures against fake information but in all speech research. Xin Wang developed a synthesis method called the "neural source-filter (NSF) model" (Figure 3).

Wang We developed a new network that synthesizes human speech waveforms by using machine learning. To synthesize speech from text, the text is first scanned and then the required

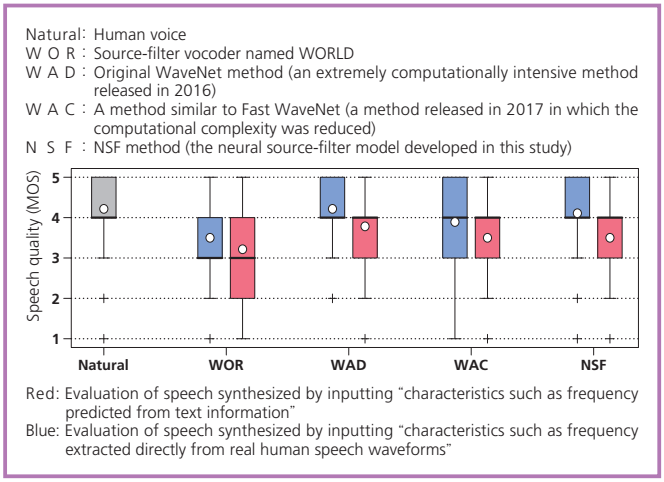


Figure 3 Results of evaluating synthesized speech using MOS method

sound characteristics are extracted, to convert the text into speech waveforms. However, until now, it was difficult to build a waveform model capable of producing high-quality speech.

Our model approximates the behavior of the vocal tract by using multiple neural networks. At first, this creates a sound like air emitted from the vocal cords, but when passed through the vocal tract networks several times, high-quality speech results.

The calculation is iterative, so computational complexity increases, but the computation time is much shorter than with existing methods. To synthesize one second of speech, it is necessary to calculate 16,000 points of data that make up the waveforms. Our method allows parallel processing, so if there are 100 processors, each processor can calculate 160 points.

Yamagishi A speech synthesis method developed by a subsidiary of Google has been regarded as high quality until now. However, it calculates the 16,000 points in chronological order, so it takes a considerable amount of time. Using that method, 100 points are calculated in one second, but in the same time, 2 million points can be calculated using Xin Wang's method.

—— So, you developed the new model and algorithm in direct competition with a subsidiary of Google?

Yamagishi That's right. Xin Wang enjoys thinking through difficult machine learning algorithms. The model he came up with originally was pretty complex, so I suggested that he eliminate certain parts, but he chose to do something different. It turned out that he was right.

Protecting in advance the information that will be targeted next

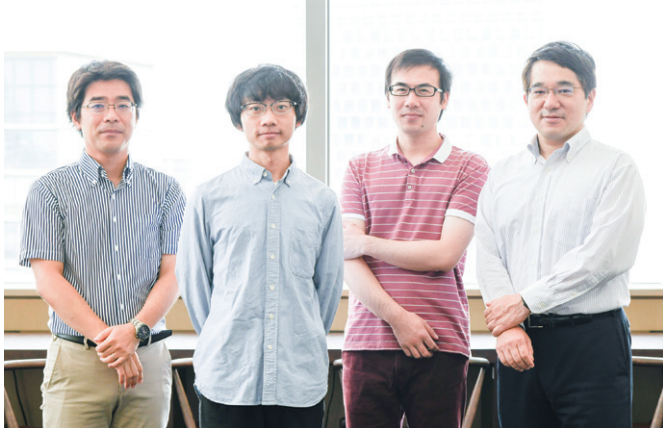
—— How will you move forward with your research?

Echizen We will continue to research protective measures while also spending time anticipating threats that may occur in the future and communicating those threats. Our focus will be on images, speech, text, and perhaps haptics. When fake information that combines these and converts them to fit a targeted person's likes and thoughts appears, how will we detect whether this information is malicious? We must consider a new threat model.

Fang With regard to text, I am beginning to research fake reviews. On websites such as shopping sites, users write reviews and many people read these reviews when they choose products. However, if a malicious organization were to use a computer to generate and post a huge number of negative reviews for a particular product, it would present a serious problem.

We verified whether it is possible for a computer to make such a judgment by combining two machine learning networks. First, we read real posts into a computer, and if the contents of the post were negative, we had the computer predict how it might continue. We then input the prediction results into another network and had it judge whether the content was negative. This allowed us to churn out negative texts with different wording.

When we asked people to read one real and three fake reviews and decide which was the real one, the rate of correct answers was only 20–30%. Since random guesswork would have a rate of 25%, this shows that people had difficulty determining the authenticity of the reviews. Research on protective measures is ongoing, but for example, one method is to consider ways of detecting whether a post is computer generated without looking at the contents of the post. Another method is to check the contents of the post using natural language analysis. In either



case, text has few of the traces of artificiality found in images, so it is not easy.

Wang As I mentioned earlier, the problem of fake speech is becoming increasingly important. I think there are two points that should be dealt with in our research. The first is constructing a general system to detect fake speech. This year's ASVspoof Challenge demonstrated that high-quality synthesized speech can be detected with high accuracy, but it requires a detection system that merges multiple subsystems with different configurations. If a single subsystem is used, the detection accuracy declines substantially. The second point, as one of the organizers of the ASVspoof Challenge, is to improve the databases containing new speech synthesis technology.

Yamagishi When we attempt to determine the authenticity of contents, including images, speech, and text, there is overlap with discussions about fairness and explainability of systems. The systems we habitually use give us various recommendations, but questions arise, such as whether someone is applying some sort of bias in the background or how to explain that point. I think that this is a topic that many people should be discussing.

(Photography by Yusuke Sato)

Note

[1] Automatic Speaker Verification Spoofing and Countermeasures Challenge

A Word from the Interviewer

Research on computer generation of images and speech is exciting and should be enjoyable. However, it is unfortunately true that researchers must also think about methods of attack using fake information and protective measures. The Echizen-Yamagishi team is crossing over the boundaries between labs, organizations, and countries to tackle this difficult problem with researchers from around the world. At the end of the interview, I asked Xin Wang and Fuming Fang what they thought about the research environment at NII. When I did so, Professor Echizen and Professor Yamagishi, perhaps aware that it might be difficult for the pair to reply with their supervisors present, tactfully left the room, saying that they had other things to do. Without this degree of flexibility, perhaps it would not be possible to conduct research across boundaries. The pair replied as follows: "It's great that there are links with laboratories around the world. In 2018, I had the opportunity to do research overseas" (Xin Wang). "Research is what I want to do more than anything else, and the environment here allows us to focus on research" (Fuming Fang).

Nobuyuki Yajima

Senior Researcher, Nikkei BP Intelligence Group,
Nikkei Business Publications, Inc.
Studied mathematics at university with the aim of becoming a computer engineer, but joined Nikkei McGraw-Hill, Inc. (now, Nikkei Business Publications, Inc.) in 1985 and became a reporter for the *Nikkei Computer* magazine. Worked as editor-in-chief of *Nikkei Computer* from 2009. Assumed his present post in 2016.

Evaluating the Authenticity of Internet Discourse

Efforts by FactCheck Initiative Japan, and the significance of the research

Kentaro Inui

Professor, Graduate School of Information Sciences, Tohoku University / Visiting Professor, NII

Interviewer: Jun-ichi Taki

Senior Writer, NIKKEI INC.

Professor Kentaro Inui of Tohoku University has been using natural language processing technology to determine the authenticity of statements circulating on the Internet. He has participated in the activities of a nonprofit organization, FactCheck Initiative Japan (FIJ, Chairman: Shiro Segawa), for several years and has expanded his activities in collaboration with journalists. He has stated, "Verifying the authenticity of rapidly spreading information on the Internet is a major concern in informatics."

About FactCheck Initiative Japan

—— What kind of organization is FactCheck Initiative Japan?

Inui It is a group that was founded primarily by Hitofumi Yanai from the Watchdog for Accuracy in News-reporting, Japan (WANJ), which runs the media watchdog website GoHoo, and journalist Yoichiro Tateiwa. It supports fact-checking that investigates whether the news, information, and discourse widespread in society is based on truth. From the information and statements circulating on social networking services (SNS), FIJ collects those where there is doubt that they are based on fact and gives them a score indicating the degree of doubt. It then provides this score to media outlets and other organizations involved in fact-checking. FIJ doesn't do the checking itself.

At my laboratory, we invented and improved a computer program that executes the work of locating questionable information, scoring it, and providing the score to the media. Engineers from SmartNews, Inc., which supports FIJ, implemented the program. Previously, people were looking through tens of thousands of items of information manually, and they were spending about ten hours a day searching for questionable information. By computerizing the task, it can now be done within an hour. However, the computer's role ends there, and it can't judge whether the information is true or false. In the end, that's the job of the journalist.

—— Why did you become involved in fact-checking?

Inui Misinformation and false rumors have always existed, but in the Internet age, it has become easier for individuals to spread information, and erroneous information spreads rapidly. This is a major concern from the perspective of information science.

The Great East Japan Earthquake struck a year after I came to Tohoku University, and groundless information such as "mouthwash protects against radiation exposure" circulated. In response, both posts approving of the information and posts denying or raising doubts about the information often appeared. I attempted to detect these and make them available side by side because I thought it would be good if people could read both the affirmative and negative posts and judge whether the information is true. I also analyzed how posts on particular topics spread and converge.

Actually, even before coming to Tohoku University, I participated in a project on evaluating the authenticity of information by the National Institute of Information and Communications Technology

Kentaro Inui

Completed a doctoral course at the Graduate School of Information Science and Engineering, Tokyo Institute of Technology in 1995. PhD (Engineering). After working as a research associate at the same university, he was engaged as associate professor at Kyushu Institute of Technology and associate professor at Nara Institute of Science and Technology, before assuming his current role in 2010. Specializes in intelligent informatics and natural language processing. Engages in fundamental research on automatic editing of linguistic information and knowledge by computers, and the supporting artificial intelligence.

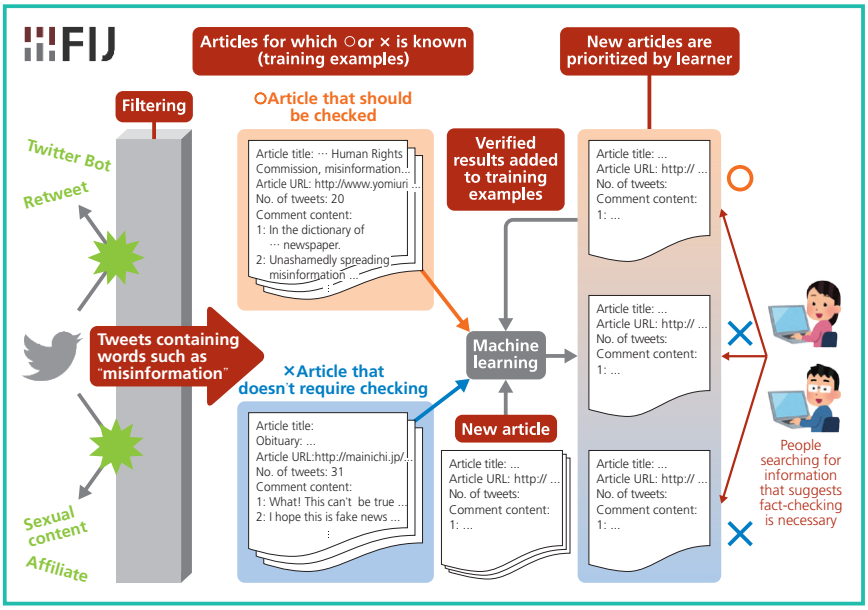


Figure 1 Narrowing down articles using natural language processing/machine learning

(NICT), led by Professor Sadao Kurohashi from Kyoto University. I worked on evaluating the authenticity of information on the Internet by using natural language processing technology. These activities caught the notice of Hitofumi Yanai and others, and I was asked to cooperate with FactCheck Initiative Japan.

Provision of questionable information was useful in elections

—— It appears that FIJ's activities have intensified as a result of elections that have drawn a lot of interest from the general public.

Inui Fact-checking is being done on a daily basis, but I don't think that the wider public is aware of it yet. FIJ considers elections to be good opportunities to make people aware of the importance of fact-checking, and it handles fact-checking of statements made during elections as one type of project.

So far, there has been the Okinawa gubernatorial election in 2018 and the House of Councilors election in 2019. In the former, approximately 5,200 reports and statements were checked by computer, and around 100 were provided to the media as candidates for checking, which amounted to 13 articles. In the latter, out of approximately 70,000 pieces of information, 72 were provided to the media, which came from 10 articles. Erroneous information was also found on all sides of the election campaigns, regardless of political party.

Significance of using real-world linguistic information

—— What significance does participating in FIJ's activities have for your research?

Inui Unlike information that is prepared for use in an experiment, real-world information sometimes goes beyond the assumptions of researchers. For example, quite often people will post a correction later in response to their own post saying, "This was wrong," but this is not a questioning or denial of the original information by another person. This kind of post must be distinguished as well.

There are various challenges, but I think that strengthening the technology using complex and varied real-world information is necessary, so that the technology becomes truly useful in society.

—— What kinds of language processing methods are used in fact-checking? Also, in the future, will computers become capable of judging the authenticity of information?

Inui The technology currently used for fact-checking is not necessarily at the cutting edge of research. Research applying deep learning to natural language processing is progressing rapidly. It is not as advanced as in the field of visual information, but we are entering a very significant era in which, for example, highly practical automatic translation machines have been emerging.

Verifying whether questionable information on the Internet is actually false information is not an easy task for a computer. This is because the truth is in the real world, and only a fraction of that appears on the Internet.

Eventually, I would like to get to the heart of that matter, but first I want to tackle determining authenticity by comparison with information on the Internet. This means locating the correct information on the Internet and comparing it against the questionable statement. Even this is not easy but considering the technological progress that has been made in the last few years, I think that it may be the next target.

—— Detecting false statements is more difficult than detecting fake images, isn't it?

Inui The technology for creating fake images is becoming more sophisticated, but evidence that a computer was involved in creating the image is left behind. This evidence may be difficult for humans to discern, but a computer can find it. On the other hand, it is difficult for a current computer to write texts that can deceive a lot of people, and these are assumed to be written by human beings. The phase of development of fake information is different in visual information and linguistic information. At this stage, careful fact-checking work by journalists is needed to detect such false statements.

(Photography by Mito Takahashi)

A Word from the Interviewer

As an example of progress in natural language processing research, I was shown AI for writing assistance offered by a venture company, Langsmith Inc., set up by students at Professor Inui's laboratory. I typed in a clumsy English sentence, and it suggested a series of succinct expressions that sounded more natural. It was amazing!

The computer does not understand the meaning of the sentence. However, using information found behind the string of words, it creates a string of words with the same meaning. I know it is merely a pretense, but it appears as though the computer understands the meaning. I got the sense that deep learning will transform the world of language.

Jun-ichi Taki

Senior Writer, NIKKEI INC.
After graduating from the School of Political Science and Economics at Waseda University, joined Nikkei, Inc. After working in branch offices and covering corporate news, began covering science and technology, as well as environmental fields, starting from the mid1980s. Authored "*Eco-Uma ni Nore!*" (Shogakukan) and co-authored "*Kansensho Retto*" (NIKKEI, Inc.), among others.

Reflection on Research Misconduct Leads to Development of Original Science-Misconduct Prevention System

Deposition and storage of all data related to research papers

Yukihide Tomari
Deputy Director, Institute for Quantitative Biosciences, The University of Tokyo / Professor, Laboratory of RNA Function, Institute for Quantitative Biosciences, The University of Tokyo

Takashi Sutani
Lecturer, Laboratory of Genome Structure and Function, Institute for Quantitative Biosciences, The University of Tokyo

Yusuke Komiyama
Assistant Professor, Digital Content and Media Sciences Research Division, NII / Research Center for Open Science and Data Platform (RCOS), NII

A center for fundamental biological research, the Institute for Quantitative Biosciences (IQB) at the University of Tokyo consists of 17 laboratories and approximately 200 faculty and students. The impetus for its establishment was two cases of research misconduct at its predecessor institute. I asked Professor Yukihide Tomari and Lecturer Takashi Sutani from IQB, and Assistant Professor Yusuke Komiyama from NII about initiatives to prevent science-misconduct.

Overcoming research misconduct

—— The background to IQB's establishment

Tomari The Institute for Quantitative Biosciences (IQB) at the University of Tokyo was established in April 2018 in a reorganization of the Institute of Molecular and Cellular Biosciences (IMCB). The impetus for this reorganization was two cases of research misconduct at IMCB that came to light in 2014 and 2017.

In conventional biological research, there were many studies with poor reproducibility based on faulty methodology, and this was one of the factors that lead to the misconduct. Reflecting on this, the current research institute aims to advance bioscience research with an emphasis on "quantitativeness" and "reproducibility" by actively adopting new methods that have excellent accuracy, resolution, and completeness, combined with a data-driven approach. Therefore, the word "quantitative" was included in the name of the institute. There are several bioscience research institutes overseas that have the word "quantitative" in their name, but IQB is the only one in Japan at present.

—— Initiatives to prevent research misconduct

Sutani With the new start, we proceeded to establish a data management system that would not allow misconduct to happen. The system development project started in August 2017, prior to the opening of the institute.

Research in a wide range of fields is carried out at the institute, including molecular biology, biochemistry, cell biology, and structural biology. However, regardless of the field, the main point of the science-misconduct prevention measures is "proper storage and management of raw data on which research papers are based." By comparing the raw data and the research paper, major research misconduct such as fabrication and falsification naturally comes to light.

What is important in data management is "a system that allows immediate access to the research data when there is a suspicion of misconduct." In the case of researches based on Grant-in-Aid for Scientific Research (KAKENHI), there is an obligation to store and disclose the data, but if the data are simply stored on an individual's PC or in a laboratory's storage, they cannot be found immediately and there are risks such as data disappearing due to a hardware fault.

Therefore, we built our own "Manu-

Yukihide Tomari

Completed a doctoral course at the Department of Chemistry and Biotechnology, Graduate School of Engineering, The University of Tokyo in 2003. PhD (Engineering). After working as a postdoctoral researcher at the University of Massachusetts, USA, became a lecturer at the Institute of Molecular and Cellular Biosciences, The University of Tokyo in 2006. Became an associate professor at the same institute in 2009 and a professor in 2013. Has served as deputy director since 2017. With the reorganization of the institute, became deputy director/professor at the Institute for Quantitative Biosciences, The University of Tokyo in 2018. Conducts research on the molecular mechanisms of small RNAs.

Takashi Sutani

Completed a doctoral course at the Division of Biological Sciences, Graduate School of Science, Kyoto University in 1999. PhD (Science). After working as a postdoctoral researcher at Harvard Medical School, USA, and project assistant professor at Tokyo Institute of Technology, became an assistant professor at the Institute of Molecular and Cellular Biosciences, The University of Tokyo in 2010 and a lecturer in 2015. With the reorganization of the institute, became a lecturer at the Institute for Quantitative Biosciences, The University of Tokyo in 2018. Conducts research on control of the higher-order structure of chromosomes.

script scan & Original data Deposition (MOD) system" that automates the task of uploading research data and figures to storage in the cloud after a paper is accepted. Using Google's cloud service "G Suite for Education," we coordinated Drive, Forms, and Spreadsheets functions with Google Apps Script. We also created a system for checking that the uploaded figures do not show evidence of fraudulent manipulation.

Building a system to store and manage research paper data

—— Types of data subject to management

Tomari The data on which a research paper is based are collected, including all raw data outputted from, for example, measuring instruments and intermediate processed data obtained by processing the raw data. Also, the accepted manuscript and figures and a checklist showing that the paper was written properly must be submitted.

When building the system, we began by deciding upon the criteria for the types and quality of data that should be stored. As well as image data obtained from microscopes and gel electrophoresis, output files specific to various analysis equipment used in biosciences are stored as raw data. Depending on the study, there may be as much as 40 gigabytes of raw data per paper.

In the case of intermediate processed data, the names of software used for data processing, such as image adjustment and statistical analysis processing, are also recorded. For figures, the data before rasterization, i.e., with layer information remaining, are stored.

The paper's authors are required to promptly report that the paper has been accepted via a web form, and then to upload the manuscript and figures within three days and the raw data and intermediate processed files within one month of the paper being accepted. The paper's authors save different types of files in folders with a defined hierarchical structure. Our system detects the hierarchical structure on the system side. Also the proposal method uploads the research files into the hierarchical structures.

Stored figures are checked by the Office for Research and Ethics Promotion within the institute to find out whether there is any evidence of image manipulation that raises a suspicion of misconduct. The system uses our original image filters to bring to light evidence of processing such as copying & pasting. Full-time members of staff who are experts in science-misconduct detection are engaged in the detection work.

The system started operating at almost the same time as the institute was founded (January 2018), and it is currently functioning extremely well.

Towards a universal RDM service that can be used by other institutes

—— Expanding the science-misconduct prevention system to other research institutes

Komiyama NII and IQB want to work together to make a similar science-misconduct prevention system. The system is available to universities and research institutions nationwide on a universal research data management (RDM) service.

NII is currently building the NII Research Data Cloud (NII RDC) (https://rcos.nii.ac.jp/service/). This infrastructure platform allows individual researchers and research groups to manage, publish, and discover research data and related information. R&D of NII RDC started in April 2017 under Center Director Kazutsuna

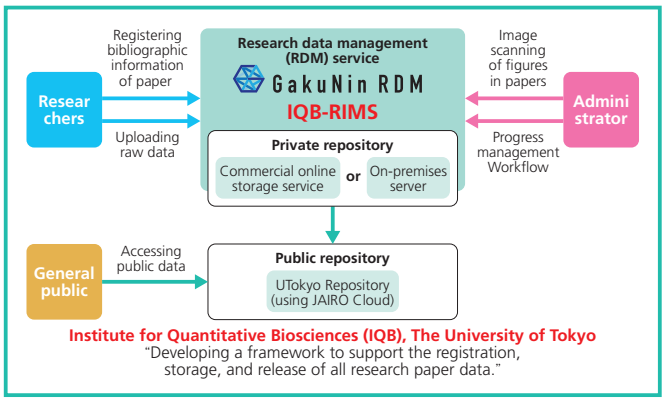


Figure 1 The Research Integrity Management System (IQB-RIMS) built by the University of Tokyo's Institute for Quantitative Biosciences by using NII's research data platform, NII RDC.

Yamaji of the Research Center for Open Science and Data Platform (RCOS), NII. NII. We aim to demonstrate the RDM service at academic institutions nationwide was conducted from April 2019. We schedule NII RDC for full-scale operation in the second half of 2020.

NII RDC consists of multiple services with different roles. Among these services, GakuNin RDM (https://rdm.nii.ac.jp), RDM service, allows on-premises servers owned by institutions and cloud services to be used as GakuNin RDM storage, as well as offering functions for version management of uploaded files and storage of research trails.

Tomari The existing MOD system hard-codes[1] paths and folder names based on IQB's IT environment. If we modify it to use more generic codes and allow data to be uploaded to GakuNin RDM, it will be possible for any research institution to use the system.

—— The future direction of system development

Komiyama Currently, NII and IQB are jointly developing the Research Integrity Management System (IQB-RIMS) in the form of a GakuNin RDM plug-in function. Subsequently, we hope to evolve GakuNin RDM into a service that is used to prevent misconduct and capable of supporting research activities. GakuNin RDM is already equipped with functions for communication between researchers, but the overall goal that we are aiming at with NII RDC is to provide a service that allows researchers to share and publish research data by registering the data in an institutional repository.

Sutani With the existing MOD system, research data are compiled and uploaded after a paper has been accepted, but the ideal situation would be for data to be uploaded at the time of each experiment. I would like to investigate in that direction in the future.

Komiyama That's interesting. It would mean that experimental data that did not yield the intended results—failed data—would also be uploaded. Looking at that data from a different perspective and re-using it could be one form of "open science." In fact, journals that publish negative results have appeared recently. If the idea that "failure is a form of success" spreads, falsifying experimental results may start to make less sense.

(Interview/Report by Naoki Asakawa[Deputy Editor-in-Chief, Nikkei xTECH/Nikkei Computer], Photography by Yusuke Sato)

Note

[1]Hard-coding: Embedding a specific operating environment directly into the source code when developing software.

Yusuke Komiyama

NEWS 1

"Programming is fun!"

Computer Science Park in Kasumigaseki

As part of "Children's Day for Visiting Kasumigaseki," NII held the "Computer Science Park in Kasumigaseki" at the Ministry of Education, Culture, Sports, Science and Technology (MEXT) on August 7 and 8. During this event, MEXT and other government ministries join together to provide ministry tours, introductions to their work, and participatory programs. The aim is to create an opportunity for children to learn broadly about society during the summer vacation through hands-on activities, and to deepen understanding of government policies.

The concept of the Computer Science Park is "a playground where children can learn 'programming thinking' without using computers." The instructors included NII's Professor Emeritus Kenichi Miura.

The children learned about iteration by thinking about multiplication using "MultipliCubes," invented by Professor Emeritus Kenichi Miura, and they learned about the concept of programming while having fun using coding blocks to create a program that would make a robot move. They also tried programming a character on screen by inputting codes into a computer. When the character moved in the way that the children had programmed it to, there were happy shouts of "I did it!" "It moved properly!"

[Top] Children thinking about multiplication using "MultipliCubes" [Right] Having a go at programming a robot using coding blocks

NEWS 2

Start of 2019 Public Lectures "The Forefront of Informatics"

First lecture on "matching theory," second lecture on "3D sensing"

The 2019 Public Lectures "The Forefront of Informatics" have started. In these lectures, NII researchers explain cutting-edge research in the field of informatics to the general public in an easy-to-understand manner. This year there will be four lectures on various themes including computer vision and theoretical computer science.

The first lecture was held on July 2 (Tue). NII's Assistant Professor Yu Yokoi of the Principles of Informatics Research Division gave a talk titled "Making Everyone Happy!? The Mathematics and Computation of Matching — How to Determine Smart Allocations —."

As a topic in which both theory and application are evolving, she introduced bipartite matching models that match the members of two groups. Assistant Professor Yokoi explained the Gale–Shapley algorithm, which makes it possible to find stable matchings (highly fair matchings capable of satisfying everyone) whatever the order of preference of each participant. This algorithm and its improved versions are actually used in Japan and the USA in applications such as assigning resident physicians to hospitals and students to laboratories. By introducing and improving the algorithm, the match rate and participant satisfaction increase.

[Left] Assistant Professor Yu Yokoi
[Right] Assistant Professor Satoshi Ikehata

However, in the real world, there are various constraints, such as wanting to deploy physicians in a balanced way that takes account of their specializations, wanting to avoid concentrating physicians in urban areas, and wanting to set lower limits for the number of people assigned, and there are many problems that cannot be dealt with using the existing theory or model. Therefore, Assistant Professor Yokoi is working to solve these constrained problems by developing the existing matching theory. She explained her research results, which show that when constraints have a certain good property, matching can be calculated quickly, and that if a stable matching does not exist, a highly fair matching can be made possible by relaxing the stability condition.

The second lecture was held on September 10. Assistant Professor Satoshi Ikehata of the Digital Content and Media Sciences Research Division talked about the basics of 3D sensing technology, familiar examples of its application, and cutting-edge 3D sensing methods combined with deep learning.

3D sensing technology, such as 3D maps, 3D real-estate property information, and games using AR/VR, has become familiar to us in recent years. Assistant Professor Ikehata explained that 3D sensing technology uses sensors and information processing techniques to perceive the world in 3D like a human being does.

As an example of his research, Assistant Professor Ikehata introduced "stereo matching," a method of estimating parallax by matching various parts of two images taken by two cameras placed horizontally, one on the right and one on the left. He said that application to autonomous driving and robots is accelerating as stereo matching shifts from local pattern matching that looks only at each image pixel to global optimization that estimates the parallax of all pixels simultaneously, as well as shifting to data-driven deep learning.

Assistant Professor Ikehata also introduced "photometric stereo," which uses one camera to photograph an object under light coming from different angles and then performs 3D reconstruction based on multiple shading patterns for a single object. He explained some of his own research results, such as photometric stereo for general materials, and a photometric stereo method that involves integrating all shaded images into one image and using that integrated image for deep learning.

NEWS 3

Japanese Culture and AI Symposium 2019 —AI for Reading *Kuzushiji* is Now Ready!

To be held on November 11—registration now open

The "Japanese Culture and AI Symposium 2019—AI for Reading *Kuzushiji* is Now Ready!" (admission free) will be held on November 11 at Hitotsubashi Hall (Chiyoda-ku, Tokyo). It will be hosted by the Research Organization of Information and Systems' Center for Open Data in the Humanities of the Joint Support Center for Data Science Research (ROIS-DS), along with the organization's National Institute of Informatics and the National Institutes for the Humanities' National Institute of Japanese Literature. This symposium will introduce the forefront of *kuzushiji* (cursive characters) research worldwide, as well as discussing the past, present, and future of research using AI to decipher *kuzushiji*.

A competition called "*Kuzushiji* Character Recognition: Opening the Door to A Thousand Years of Japanese Literate Culture" is being hosted on Kaggle, a platform that attracts worldwide attention for its many AI competitions. This symposium will bring together the winners of the competition, and they will explain their *kuzushiji* recognition algorithms. Also, leading researchers working on writing culture in Japan, from *mokkan* (inscribed wooden tablets) to *kuzushiji*, have been invited, and they will give lectures and demonstrations introducing the cutting edge of research.

This is a unique opportunity to experience how far the technology for deciphering *kuzushiji* has come and its potential. We invite you to join us. Registration is via the following website.

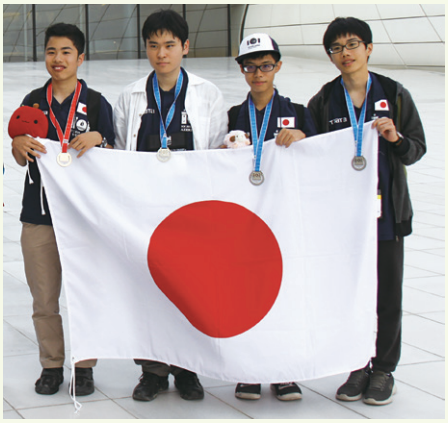
http://codh.rois.ac.jp/symposium/japanese-culture-ai-2019/

* See "Hey, this is great!"

NEWS 4

Projection celebrates Japan's representatives at the International Olympiad in Informatics

[Left] Projection shown on building facade
[Bottom] (From left) Masataka Yoneda, Koichi Namekata, Yasutaka Hiraki, and Sora Todaka wearing their medals and smiling for the camera [Photo provided by Japanese Committee for International Olympiad in Informatics]



September 5 to 20, a projection celebrating the performance of Japan's representatives at the 31st International Olympiad in Informatics (IOI) was shown on the facade of the National Center of Sciences, which houses NII. NII has actively endorsed the IOI by featuring the event in NII Today (https://www.nii.ac.jp/about/publication/today/76.html) and holding a workshop entitled "Aim to be a Future Informatics Olympiad Medalist!" where junior and senior high school students took on the challenge of solving problems for the IOI.

This year's event was held in Azerbaijan during August 4 to 11, and four senior high school students representing Japan won one gold and three silver medals.

The gold medal was won by Masataka Yoneda (2nd-year student at Senior High School at Komaba, University of Tsukuba), and the silver medals were won by Sora Todaka (3rd-year student at Miyazaki Nishi High School), Koichi Namekata (3rd-year student at Senior High School at Komaba, University of Tsukuba), and Yasutaka Hiraki (2nd-year student at Nada High School).

The IOI is one of the International Science Olympiads aimed at students up to senior high school level. A total of 327 students from 87 countries and regions around the globe took part, and they competed on their ability to solve problems in mathematical information science, such as devising algorithms. Gold medals were awarded to participants who scored in the top one-twelfth, silver medals to the next two-twelfths, and bronze medals to the next three-twelfths.

The projection showed images of the contestants engrossed in programming and Japan's representatives smiling with their medals around their necks. Passersby stopped to look at the projection.

SNS

"Hey, this is great!" Hottest articles on Facebook and Twitter (June 2019-August 2019)

National Institute of Informatics, NII (official) Facebook
www.facebook.com/jouhouken/

[News Release]
AI researchers/engineers worldwide tackle *kuzushiji* (cursive characters) recognition: Global competition hosted on Kaggle from July
From July to October this year, a global competition titled "*Kuzushiji* Character Recognition: Opening the Door to a Thousand Years of Japanese Literate Culture" will be hosted on Kaggle, the world's largest machine learning competition platform. We can expect not only that this competition will accelerate the development of groundbreaking *kuzushiji* recognition techniques but also that the *kuzushiji* datasets will attract worldwide interest in Japanese culture. (07/10/2019)

National Institute of Informatics, NII (official) Twitter
@jouhouken

[News Release]
AI researchers/engineers worldwide tackle *kuzushiji* (cursive characters) recognition: Global competition hosted on Kaggle from July (07/10/2019)

Bit on Twitter! Twitter
@NII_Bit

The final NII Research 100! Thank you to MCs Professor Hideaki Takeda (@takechan 2000), Associate Professor Ikki Ohmukai (@i2k), and Ayaka Ikezawa (@ikeay)! (06/01/2019)

* Some text edited/omitted.

Essay

Protect SNS from information manipulation

Takayuki Mizuno
Associate Professor, Information and Society Research Division, National Institute of Informatics
Associate Professor, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies

The large-scale ongoing demonstrations in Hong Kong have taken on aspects of information warfare. Both pro-establishment and anti-establishment camps are intent on shaping public opinion inside and outside Hong Kong by feeding convenient information to SNS (social networking sites).

Recently, there has been lively debate in computational social science about the danger of information-sharing mechanisms using SNS. People unconsciously believe information that is convenient to them, and they try to increase their circle of friends by sharing this information. When they do so, they start a chain in which their friends respond by posting related information and they respond again. This is described as an "echo chamber," and it explains why opinions quickly become divided on social networking sites. People prefer sensational information, and if that information is convenient to them, they tend not to care whether it is fake news.

Professor Gary King of Harvard University revealed in an article in the American Political Science Review (APSR) in 2017 that anonymous accounts belonging to the Chinese government are publishing posts that suppress collective action such as demonstrations. These posts guide public opinion by skillfully directing people's interest to other topics, rather than by defending the government's position. We are engaged day after day in research and development aimed at garnering the collective wisdom of humanity by creating freer and more efficient SNS, based on the belief that human nature is fundamentally good. However, it appears that the technology is being used in an underhand way.

Technology for creating safe and secure SNS will be necessary in the future. For example, the blockchain method is effective for maintaining sender information and preventing falsification. This is a measure directed upstream of the information source. However, measures directed downstream are also possible. These concern who the information is provided to, that is, whether information is released to Party A or Party B. Systems that consider who is downstream and recommend the safety of the destination may be effective.

We have moved from an era in which Internet posts were thought of as simply graffiti on a street corner to an era in which they have become a public good. Such an era requires SNS that are safe and secure while also remaining free and efficient. We informatics researchers will also continue to carry out research aimed at establishing SNS that serve as a public good.

Future Schedule

October 20 | Inter-University Research Institute Corporation Symposium 2019 (Exhibitor), National Museum of Emerging Science and Innovation (Miraikan). For details, go to **https://www.4kikou.org/symposium.html.**

November 07 | 3rd session of public lecture series on "The Forefront of Informatics"—"An Introduction to Theoretical Computer Science — Between Finite and Infinite: From mathematical theory to AI and autonomous driving" (Speaker: Associate Professor Ichiro Hasuo, Information Systems Architecture Science Research Division). For details and registration, go to **https://www.nii.ac.jp/event/shimin/.**

November 12–14 | 21st Library Fair & Forum (Exhibitor), Pacifico Yokohama.

December 12–14 | 2019 Annual Meeting of the Academic Exchange for Information Environment and Strategy (AXIES) (Exhibitor), Fukuoka International Congress Center.

Publication of public information magazine for children *NII Today Jr.*

NII publishes a public information magazine for children *NII Today Jr.* to convey the fun of informatics to kids.
In 2019, NII is publishing brochures and posters (Vol. 1/Vol. 2) as part of a series "Learn with NII's Info Dog Bit-kun! Algorithms" NII researchers answer questions such as "What are algorithms?" and "What are algorithms used for?" in an easy-to-understand way. There is also a fun quiz.
You can find *NII Today Jr.* at the following website: **https://www.nii.ac.jp/about/publication/today/**

Notes on cover illustration A suspicious-looking robot is making a fake image by replacing the face in the photograph so that it looks as though a different person is committing the crime. Such "lies" are becoming more and more sophisticated as AI and other information technologies advance, and they are becoming a threat to society.

Weaving Information into Knowledge
NII

National Institute of Informatics News [NII Today]
No. 71 September 2019
[This English language edition NII Today corresponds to No. 85 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 | Masaru Kitsuregawa **Editorial Supervisor** | Ichiro Satoh
Cover illustration | Toshiya Shirotani **Copy Editor** | Madoka Tainaka
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

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