ISSN 1883-1974
National Institute of Informatics News
No.34 July 2010
(This English language edition of NII Today corresponds to No.48 of the Japanese edition)

NII Today

National Institute of Informatics News

[Special feature] Language
Using language as knowledge

Computers Read Text

Creating an Age Where Anyone Can Find the Information They Truly Need
NTCIR's Information Retrieval Ideal

Information Stops the Spread of Disease
The Allure of Multi-Faceted "Language"

[Special feature]
Language

Using language as knowledge

Text (language) is an essential part of human life.
With the advancement of computing, language has passed beyond being merely a means of communication. It now possesses value in and of itself, serving as "information" or "knowledge".
We present the cutting edge of language research using computers with advanced processing capabilities.

NII Interview

Computers Read Text

Akiko Aizawa
Professor, Digital Content and Media Sciences Research Division, NII

Computers Use "Text" to Understand Reality

Ikeya Professor Aizawa, I have heard that within the body of researchers specializing in language processing, you are dedicating particular effort to the processing of actual data which contains a high incidence of errors. I'd like to take the opportunity of talking with a "data expert" such as yourself to discuss "text".

Aizawa In recent years, language processing has come to handle truly large quantities of data. My own work primarily consists of discovering what valuable data can be obtained from disparate, high volume texts. Every day, I am made acutely aware of the tremendous demands, both in terms of computing power and sheer effort, placed by actually processing information.

Ikeya How does the handling of "text" by computers differ from that of humans?

Aizawa Generally, language is thought of as a tool for human communication, but from an information processing standpoint, it becomes an object of language processing only after it is converted into electronic form and fed into a computer. One could say that information processing is the study of what, exactly, it is for a computer to read.

Ikeya For humans, reading is normally sufficient for understanding. This, however, is not the case for computer reading, I gather?

Aizawa Correct. For a computer, "reading" is, essentially, acquiring information from a text, and utilizing that information. For example, if there were a robot which could carry on a natural conversation with a person, even if it were unable to understand text in the same way as a person, it would be fair to say that it could utilize language. Language consists of give-and-take, so, for example, when we turn to a search engine to find information, that search engine is actually acquiring information from us. If we search for "Sky Tree height", it is clear that one attribute of "Sky Tree" is "height". Given millions or tens of millions of similar queries, it becomes possible for computers to acquire a considerable amount of knowledge.

Ikeya In other words, computers learn about the actual "Sky Tree" via text?

The Importance of Determining that One Thing and Another are "the Same"

Aizawa Exactly. Our normal discussions seldom consist of general descriptions, such as "dogs are smart". Instead, they include individual statements, such as "Yesterday I went to Ginza and met X." When computers collect these "facts", the virtual world created within the computer closely resemble our own. However, an accurate grasping of "facts" from a small amount of information is not something which computers excel at. On the other hand, extracting "facts" from a massive language data set, cataloguing and gathering similar items, investigating inconsistencies and paradoxes -- this is the type of work which is best suited to computer capabilities.

Ikeya In other words, taking a statistical approach in order to extract a variety

of facts ?

Aizawa　Right. I think the essence of knowledge is the determination of what things are the same. As such, the issue of whether the facets of reality determined by computers correspond to things in the real world is critical. Proper nouns, such as place names, the names of people, and the like, have concrete counterparts in the real world, and as such serve an important role as pointers tying together the world of language and the real world. Without tackling this issue, language cannot be understood. It is with regards to this that I have been involved in the NII academic content infrastructure advancement project. If successful, I believe this will make a certain contribution to the field of artificial intelligence as well.

Ikeya　A landscape where the real world, people, virtual worlds, and computers are linked via the data of "text" is beginning to unfold. Perhaps this is what makes language so interesting?

What We Can See When We Process the Vast Volumes of Text on the Web

Aizawa　I think so. Text can serve as a meter stick by which to measure society, as well as a tool to see into people's mental activities. I think what makes it interesting is that by analyzing it, we can see things such as people's knowledge and conventional societal beliefs. I refer to using text in this way as a "language sensor". We can "perceive" and measure people's values by aggregating a large volume of text.

Ikeya　While that seems fascinating, it also seems to be extraordinarily difficult.

Aizawa　It is. Understanding meaning has been a perennial problem, and only now, after collecting and working with enormous volumes of data, along the lines of tens or hundreds of millions of documents, have we reached a basic contextual understanding of general text. The issue is that challenging.

Ikeya　For example, people are now Twittering on the Web. This means that there is now a tremendous volume of text generated by individuals on the web. What kinds of discoveries will this information make possible?

Aizawa　The Web has tied people together via computers. Until now, when we discussed value, we measured it with a somewhat uniform index, like the price of products. However, people's statements include a variety of value judgments other than monetary value, such as ease of use or sense of security. Collecting this kind of information makes it possible to gain an overall view of trends, as well as discovering certain types of diversity.

Ikeya　From the perspective of the average user, it feels as if an enormous amount of individual people's day-to-day thoughts are flowing onto the Web in the form of text.

Aizawa　That's true. I think it would be fair to say that a number of methods for recording our activities have emerged, and people have taken them up, resulting in recordings of their actions and discussions. Records persist, and in 10 years or 100 years, it may be possible to stand on a street corner, and hear a long-past conversation held there. A close eye must be kept on the direction that society, and the knowledge supporting it, is taking.

Rue Ikeya
Science Communicator

Comment from the Interviewer

Professor Aizawa's research lab has recently been tackling the issue of revealing how people read, by scanning the eye movements of test subjects reading from a computer screen. There is much that remains unknown about how the actual mechanism of reading, both by computers and by people. Seen from the viewpoint of human history, people have worked with computers for but a brief time, and, says Professor Aizawa, "with today's rapid development, we mustn't regret the effort of computing". How will the relationship between people and text, established when mankind first began writing, change in the future? Eyes are on Professor Aizawa's research.

Creating an Age Where Anyone Can Find the Information They Truly Need

NTCIR's Information Retrieval Ideal

With the spread of sites such as Google and Yahoo!,
using search engines to find information has become commonplace.
What is the optimal information for users?
We asked 3 researchers involved in NTCIR, a workshop-style project
whose objective is the continued development of information access technology.

What is NTCIR?

With the spread of computers and the Internet, hearing news reports of topics such as people playing shogi or chess against computers has become commonplace. We live now in an era where computers appear on game shows. A research and development project at IBM is currently working on a computer system which will participate as a contestant on the popular American quiz show "Jeopardy!". It contains a question response system that combines information retrieval and language processing technologies, and research is underway on question answering technologies that can also handle so-called "trick questions".

NTCIR is an internationally active, workshop style project whose objective is the advancement of these kinds of information access technologies. The expression "information access" was chosen because NTCIR's objective is a system for "supporting the users to create of new value from massive amounts of information"by retrieving relevant information for the users and supporting the users to utilize the information in the documents. Therefor NTCIR is researching information retrieval, technologies for supporting the users to utilize the information in documents such as question answering, summarization, opinion analysis, trend analysis, etc. and search appropriate questions.

The project, started in 1997, subsumes several research divisions related to information access technologies, and each division is operated by a group of researchers, who function as "organizers". The importance the project places on the workshop-style approach can also be seen in research division selection. The project's management does not unilaterally establish research divisions. Instead, researchers in related fields collect research proposals, and decisions are made after a committee considers them from the perspectives of their content, feasibility, international research trends and technological trends, and social value. Each activity cycle lasts a year and a half, with NTCIR-8 (its 8th cycle) currently underway. A single cycle's process is as follows. First, organizers propose research division objectives and evaluation methodology, discussing them with researchers who wish to participate, and determining final evaluation methods and data. After this, organizers distribute shared document datasets and query datasets. Participants use these datasets in their search experiments, thereby performing verification of the systems they have developed. These results are collected, evaluated by human assessors, and correct answers are created. In some cases, correct answer candidates, created in advance, can be used, in which case many participants evaluate and verify them, increasing the reliability and validity of the correct answer proposals. Last, the verification and evaluation results are gathered in the form of research papers, which report the achievements of the research, bringing the cycle to a close. The document data, query data, and correct answers are referred to collectively as a "test collection". These are, of course, repeatedly used by researchers participating in NTCIR for further research, but they are also made publicly available to non-NTCIR participants, so that a wider research community can test their methodologies against a validated standard. This contributed to the efficient advancement of research activities. Verifying the effectiveness of information access technologies requires, for experimentation, a large number of users and questions. However, during initial stages of research, where verification is needed every time a new idea emerges, it is difficult to gather a large number of users and perform long-term testing. By using test collections, the effectiveness of research ideas can be verified immediately, and repeatedly, through research lab experimentation, rapidly accelerating research progress.

A Half Century of Information Retrieval System Progress

NII's Professor Noriko Kando has been deeply involved in NTCIR activities since it was first established. "Computer system based information retrieval research first started in the 1950s," she said, providing the following overview of the history of the research.

Shortly after research began into information retrieval, it diverged into two streams: one focused on making it commercially viable, the other on theoretical research into search algorithms and the like. The searches

Noriko Kando
Professor,
Information and
Society Research Division, NII

Atsushi Fujii
Associate Professor,
Tokyo Institute of Technology

Koichi Takeda
Senior Manager,
Senior Technical Staff Member,
Information & Interaction,
IBM Research - Tokyo

performed on the commercial systems of the time were exact match searches, searching for exact matches of search queries. Most were simple searches that looked for matches in the titles or abstracts of papers. However, with improvements in hardware development technology came a corresponding dramatic increase in the amount of data handled, requiring a re-evaluation of search approaches. Specifically, searching for a single word would result in an enormous number of hits, while searching for multiple terms would result in excessive narrowing of search focus often producing no search results. In order to resolve this problem, the ball was passed to the other stream of research, that focused on the development of search algorithms. "While commercial systems used exact matches, their research used the best match approach. The objective of this method was to present information in response to user information needs in ranked order of relevance."

In order to put the results of search algorithm research into practice, comparative validation of technologies and methods using large scale test collections was essential. This was the start of research by experimental evaluation workshops. The first was TREC, established in the US in 1992, followed by Japan's NTCIR and then Europe's CLEF. Through the establishment of these research projects, continued research result technology transfer, and close-knit research exchange, evaluation workshops became sites for new and ever challenging research task proposals.

The Best Match Search Approach

What kind of search methodology is employed by the "best match" approach, which searches for relevant information? With the best match method, search systems rank search results in the order predicted to be most relevant to the user. It includes evaluations of the frequency with which words or strings contained in searched documents and questions appear, in what patterns they appear, document lengths, and the like, and in order to calculate their similarity, create mathematical models and "retrieval models" (vector space models, probabilistic models, language models, etc.). A variety of other heuristics are also used, such as web links and usage histories, including click logs. These are used to determine similarity between searched documents and queries, to calculate importance measures, and to rank search results. The best match method does not merely search for strings contained in the query, but for the information that matches the information needs and intentions behind the query.

The Uses of Information Retrieval Systems, Extending Beyond the Bounds of the Web

The term "information retrieval" might call to mind search engines such as Google and Yahoo!, but information retrieval technology extends beyond just the web. One of NTCIR's task organizers, Koichi Takeda, an IBM Japan employee who specializes in machine translation and text mining (*) research, says: "The majority of office information is unstructured (text, image, and other), which contrasts with structured information in a database. It is said that up to 30% of a white-collar worker's time is spent searching for and analyzing information. For improving these workers' productivity, we need to develop solutions that provide information in the way that users can easily access and manage, both quantitatively and qualitatively - that is, solutions that provide relevant information."

Wide-Ranging Research, from "Patents" to "Yahoo! Answers"

NTCIR started in 1997, and is now in its 8th cycle, gradually turning into an international research project,

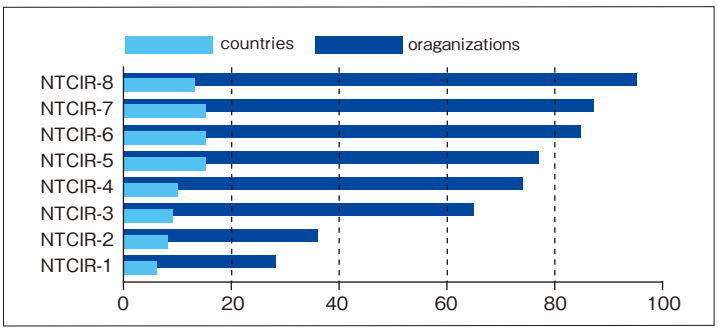


Figure 1 Trends in Countries and Organizations Participating in NTCIR

with participating researchers from 17 countries (figure 1). On the next page follows a few of NTCIR's notable research projects.

One of the fields NTCIR has researched since its formation is cross-lingual information retrieval, performing searches across multiple languages. One of NTCIR's groundbreaking researchers, who has participated in the project from its earliest days, Tokyo Institute of Technology's Associate Professor Atsushi Fujii, provides the following overview of the NTCIR's first cross-lingual information retrieval system. "Let me tell you about the cross-lingual information retrieval system developed for searching English academic papers with Japanese language queries. When performing searches with this system, there are two possible approaches: Japanese queries can be translated into English, or English papers can be translated into Japanese. The latter places a heavy burden on system processing, so generally the former is used. However, in order to obtain more accurate search results, the following approach can be used. Queries can be translated into English and searching performed. Then the top results, let's say the top 1000 results, are translated into Japanese, and a search is run again against the original Japanese query. This method came to be called "bi-directional", and is now the primary method used in cross-lingual information retrieval."

Patent information searching, started at NII in 2001, was one of the research fields that, internationally, NTCIR focused on from an early point. A number of issues were faced during research, but these were surmounted with the research assistance and backup of companies providing patent information services, and intellectual property information retrieval committee members from the Japan Intellectual Property Association. As a result, a variety of research was done on topics such as cross-lingual and cross-document type searching, and automatic "patent map" technical trend generation, and a considerable amount of the research has since been used in commercial applications.

Question answering, which extracts answers themselves from large volumes of documents, is another research division on which NTCIR has focused. Research so far diversed from working with simple factual questions such as "Who is the prime minister of Japan?", to dialog style questions with complex responses such as definitions and relationships, such as "What is information retrieval?", to answers such as "There is no answer" for unanswerable questions, and further still to cross-lingual question answering. Document searching, question answering, and summarization have developed as separate technical fields, but NTCIR believes that these technical fields will fuse in order to provide users with relevant information in an appropriate format.

Research underway on searching Yahoo! Answers is also quite unique. Yahoo! Answers is a web service where those with questions post them, and other users post answers, the best answers being marked as "Best Answer". In addition to user evaluations, NTCIR is researching a system for automatically determining the best answer. In order to increase objectivity, in addition to the best answer selected by the asker, Yahoo! Answers also had other evaluators select best answers. "Analyzing the answers selected as best answers, we saw certain trends emerging, such as comments indicating approbation of the questioner, or the insertion of URLs providing evidence for the answer. That means that evaluations differed even for the same answers based on how the responses were written. This shows that considerations regarding communication style and modes of expression were linked the type of information provision required by the asker," explains Professor Kando. These research results can contribute text retrieval an to service improvements for Yahoo! Answers, but can also be applied to a wider range of communications. For example, it would be reasonable to use such systems to automate responses to inquiries to company service desks. "Information such as that handled by Yahoo! Answers carries considerations of user privacy, making it difficult to use in research, but some of those involved in the startup of Yahoo! Answers are familiar with NTCIR's activities, and provided dataset in accordance with proper established procedures for maintaining user confidentiality."

The Allure of the Workshop-Style


NTCIR-7's research result presentation conference

NTCIR's research is supported by a large base of people. This is not limited to coordination with external organizations. Professor Kando explained the significance of coordination. "I believe that the greatest allure of NTCIR is its workshop-style, with people coming together to tackle research issues. When research is performed by a single organization, there is a limit to the number of ideas and testable approaches, and it is difficult to perform objective evaluations. During the year and a half of a project cycle, NTCIR establishes several roundtable meetings as opportunities for exchanging ideas. There is particular excitement among researchers concerning initial results reported at the meeting roundtables. Researchers are tackling the same issues, so there is a high degree of identification with details which cannot be fully encapsulated in papers, such as approaches to

the research theme, and experimental know-how."

The workshop-style also presents significant advantages in terms of technological coordination. For example, take question answering systems. These can be broken down into multiple phases, such as information collection, analysis and extraction, and aggregation and presentation. In other words, component-based is possible for systems. This means that there are many cases where, instead of a single organization developing every function, individual organizations develop the modules in which they have particular strengths, combining them to form a superior system.

Future Developments in Information Access Technology Research

NTCIR-8 will ends this June, but what will future NTCIR activities bring? "The number of information retrieval researchers is low in Japan, compared with other countries, and I believe that reinforcement is needed," says Associate Professor Fujii, who provides direction to students himself. "As someone who works in the classroom, I hope to raise not only people who can design and develop search systems, but also people who can use test collections and perform appropriate system evaluations. Evaluating search systems is just as important, and as difficult, as designing and developing them. I also think that evaluating search systems and grading students share something in common. Test collections, and test questions administered to students, must be based on fair evaluation standards, and a wide variety of questions, materials used in solving them, and correct answers, with appropriate difficulty levels, must be created. Students and systems evaluated with these must be at a level at which they can perform practical tasks in the real world. My desire is to nurture students and systems which will contribute to actual society." Dr. Takeda, as a member of the business world, expressed his visions, saying, "I would anticipate that emerging applications of NTCIR research are going to be proposed to meet real world demands. So far, research has focused on individual components such as searching, translation, or text mining. By combining them in the future, more sophisticated information access solutions will be introduced to innovate office activities." Lastly, Professor Kando discussed two future objectives. "The first is exploratory search. Web search engines are used for searching for information regarding actualities, and when the user knows that there are answers, such as searching for NII maps or tomorrow's weather forecast. However, there are also a large number of cases involving taking an interactive search and learning approach, such as when it isn't clear what the user is looking for, when the topic is in the area which is not falimilar with the user, or when the goal of the search is unclear. I hope that this kind of interactive exploratory search and information utilization is made possible in order for the answers to which search systems lead users be brought even closer to what the users are looking for. For example, consider a hypothetical mother searching for a nursery school in which to enroll her child. Parents naturally want to send their children to good nursery schools, but for a first-time parent, I don't think they know what "good" aspects to look for. I think, in this kind of case, that these aspects must be provided as possible selections. Another example of people searching very vague topics would include, for example, high school students who wish to find out about "college entrance exams". If searching for "college entrance exam" produced university rankings by department, procedures for taking entrance exams for foreign universities, graphs showing the career tracks of graduates, and the like, it would be easier for said high school students to take the next step. I want to create systems that make it possible for users to take an exploratory approach to information searching, learning, and investigation, by presenting the perspectives necessary for searches, categorizing and aggregating search result information, and arranging it for presentation. This technology, by aggregating massive volumes of data, would support the creation of new value by users, pursuing the weaving of knowledge from information, the goal of the NII, from the standpoint of information access research. Interactive information access technology evaluation methodology must be established as a foundation for this research, and research into this methodology is particularly active on an international level.

The second objective is closely related to the first, being the functioning of NTCIR as a community in the true sense of the word. I hope that the opportunity presented by NTCIR is made full use of for researchers, as organizers and participants, to openly develop and advance their own research, and to nurture students and young researchers. I believe NTCIR's role is not to lead this self-motivated community, but to provide support for it. Doing so would increase the effectiveness of the workshop-style approach, resulting in coordination which makes optimal use of individual specializations, and the creation of even better systems."

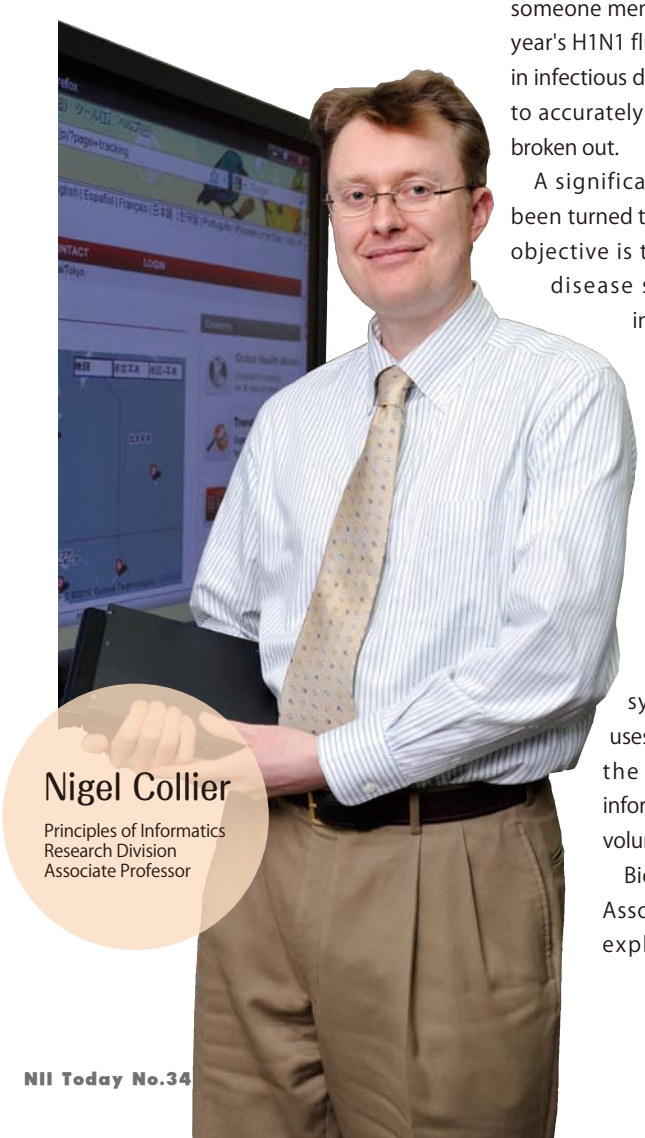
This open research style is sure to generate new, heretofore unimagined technologies and value.

(Written by Takuya Kudo)

*Text mining: includes (1) information extraction from large volumes of text data - for identifying words, named entities (names of people, places, etc.), emotional expressions, "subject/object-predicate" expressions, and the like, and (2) analysis of their frequency, patterns, and relationships in order to acquire insights that is not immediately apparent, and to support document organization and report generation.

Information Stops the Spread of Disease

**Are you familiar with the term “text mining”?
Text mining is a human language technology that takes unstructured text, analyzes the meaning of words and their interrelationships and evaluates the usefulness of information for a particular task.
BioCaster, a project that uses text mining to tackle the problem of infectious disease, is currently being carried out by research organizations in multiple countries, working together.
This document provides an overview of the leading developments in BioCaster activities.**



Nigel Collier
Principles of Informatics Research Division
Associate Professor

The Utility Value of Web Text

The first thing that comes to mind when someone mentions infectious disease is last year's H1N1 flu epidemic. What is important in infectious disease surveillance is the ability to accurately grasp where the disease has broken out.

A significant amount of attention has been turned to the BioCaster project, whose objective is the prevention of infectious disease spread through the use of information technology. A wide range of infectious disease related information on the Web is automatically collected and analyzed by the system, which then publishes warnings, countermeasures, and reference materials related to treatment on the Web. These are the main activities of the project. The system stands apart in that it uses text mining in order to extract the infectious disease related information it needs from a massive volume of data.

BioCaster’s research leader, NII's Associate Professor Nigel Collier, explains , “When the avian flu epidemic started, I thought my research in the field of natural language processing could contribute in the form of a monitoring system for infectious diseases.”

Following is an overview of the BioCaster system. First, infectious disease related information is extracted from massive volumes of natural language text on the Web. Then, key concepts, such as symptoms, viruses, dates and times, and locations, are identified, and structured events are extracted. These are ranked in order of urgency, and published on the Web, to be used by specialists in establishing disease countermeasures. The functions of each phase are described in more detail below.

Collecting Information from a Wide Range of Sources, from News Reports to Twitter

The BioCaster system’s infectious disease monitoring starts with information acquisition. It collects information from a wide variety of open media sources on the Web, from news reports and announcements by public institutions to public e-mail list discussions. One of the Web’s defining traits is that any individual can be an information source. Associate Professor Collier seized on this. “We need to think of social media, like blogs, Twitter, and Facebook, as information sources, as there are some regions whose public institutions largely lack the infra-

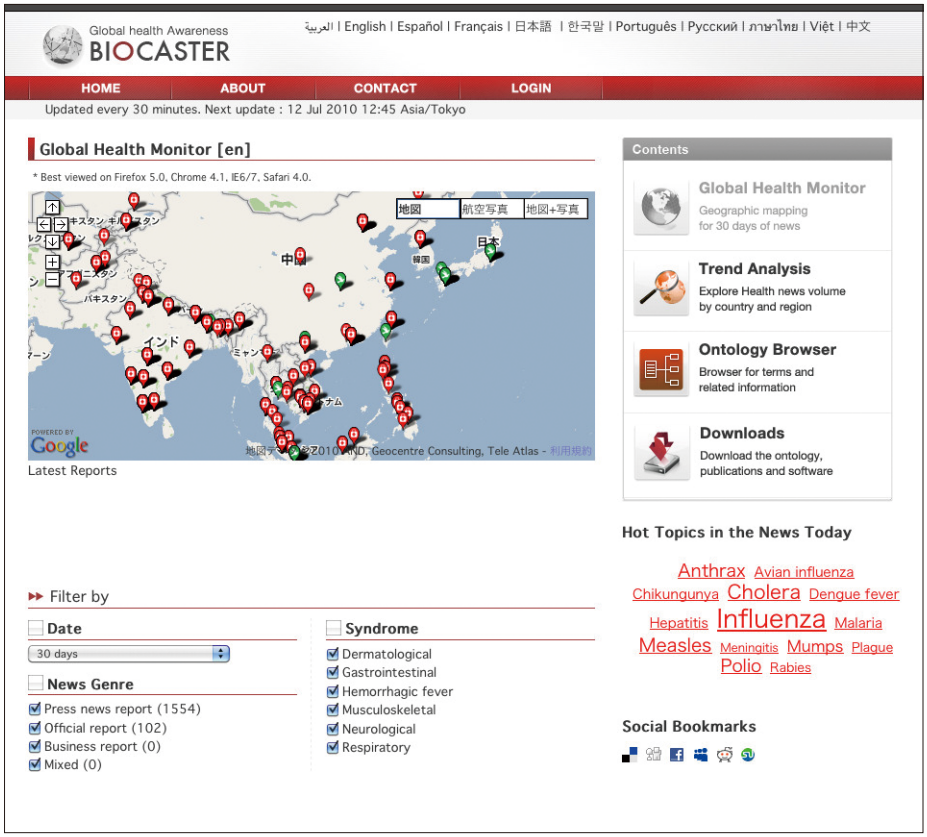
structure for providing information. Of course, since this information is coming from individuals, its credibility level may not be high. However, I think it's very valuable to design a system for extracting necessary information from those sources."

Multilingual Information Support

"The project started with English, Thai, Vietnamese, and Japanese, and is currently being extended to cover 12 different languages," says Associate Professor Collier regarding the BioCaster's multilingual diversification. The collected information in each language is translated into English within the system before categorization. This comes with its own problems, related specifically to the field of infectious disease. It is rare for an infectious disease to go by a single name. The H1N1 flu virus is a good example of that, with multiple names, such as "Swine Flu", "Swine Influenza", and "Pig Flu" in the English language alone. This frequently results in the name of a disease being translated into English with a name that English speakers do not use. This problem is resolved by registering individual word combinations in the system, but, as is imaginable, this takes considerable time and effort. This part of the process requires direct human involvement.

Timeliness Improvement Is an Issue

The information, translated into English, is categorized using text mining. The BioCaster system focuses on words which are highly related to information categories, such as disease names and the names of locations where infection is occurring. These relationships are used in identifying individual categories. Currently, from a text categorization perspective, information category identification accuracy is 70% as accurate as a human specialist, but as Associate Professor Collier says, "there are still many issues which must still be resolved." "The timeliness of the information is still an issue when it comes to preventing the spread of a disease. Text mining can only be performed when the information has been gathered together, but when a disease breaks out, the amount of information available on the Web is extremely limited. Once sufficient information has been published, the disease has already spread. It would be best to be able to detect the outbreak of an infectious disease while the amount of information available is still low, and predict the disease's spread." The categorized information is registered in a database, and can be viewed from the BioCaster website. "Our primary objective is to provide information to public health related personnel worldwide, but we would be very pleased if at the same time this information helped with infectious disease awareness among the general public."



■ The BioCaster Web portal showing the regional spread of diseases in Asia

Coordination with Organizations both Domestic and Abroad, and Future Vision

BioCaster's activities are supported by the connections between a variety of organizations. Groups such as the National Institute of Infectious Diseases, the National Institute of Genetics, Okayama University, the Vietnam National University Ho Chi Minh City, and Thai's Kasetsart University are working together to improve the system's capabilities. "Parts of our system still require human intervention. However, the system is making steady strides. We would like to continue our research in order to create a system which can be used to help prevent the spread of infectious disease by providing information collection support, making society safer."

Improvements in the BioCaster system's capabilities and usage scope will establish it as part of the world's public health information infrastructure. This will help contribute to the saving of people suffering from infectious diseases worldwide due to insufficiently developed medical and information technologies.

(Written by Takuya Kudo)

That's Collaboration:
NII-R&D Center
Young Researcher's Round-Table

The Allure of Multi-Faceted "Language"

Language - we use it every day, in our conversations, in emails, and on the web, without giving it a second thought. Researchers see in it a wealth of mysteries that spur intellectual curiosity. Three researchers gathered to discuss the allure of language, with its many facets, including written language, spoken language, and the text used on the web, as well as aspects of Japanese they are currently interested in, and more.

The Wide Scope of Language Research, from Web Searches to Sign Language

- I've heard that language research encompasses an enormous range. First, please let us know what exactly each of you is researching.

Uchiyama At the Research and Development Center for Scientific Information Resources, we put scholarly papers and Grants-in-aid for Scientific Research results in database format, releasing it publicly. My own research there is on extracting useful information from the content which has been converted into text form. Some scholarly papers use highly specialized jargon and can be rather impenetrable, while other papers are comparatively readable. Identifying the difficulty of the terminology involved makes it possible to narrow down search results to those that correspond to the reader's level. My research involves the language processing that forms the foundation of that process.

Bono I've always been interested in spoken language and gestures. For the last several years, as an extension of that interest, I have been researching Japanese sign language. One of the NII's projects is the release of The Speech Resources Consortium audio data as research material. In the future, I would like to offer sign language, which has no written form, as a language research material.

Abekawa I work at the Research and Development Center for Informatics of Association, and am developing an associative search* system that uses natural language processing technology. Currently, I'm involved in the renewal of the Webcat Plus library search system which the Center helped develop, putting on the finishing touches.

- Can you share what will make the newly renewed system special?

Abekawa Let's say, for example, that a user does a search for "natural language processing". With the technology used up to this point, this was broken down into the words "natural", "language", and "processing" in order to perform a search, making it difficult to find books about what the user was actually searching for, "natural language processing". By having the computer recognize that as a single term, the system can provide search result information that matches the user's request.

The Impetus for a Desire to Master "Language"

- I see that you cover a diverse range of research fields. What initially got you interested in your research?

Uchiyama When I was a student, I wasn't very good at English (laugh), and I once tried a machine translation system that instantly translated Japanese into English. However, the English it output to the screen was gibberish. I wondered what kind of processing the computer was performing in order to translate the entered Japanese into English - what kind of system design was used. That was my gateway into language processing research. Also, while I can accurately convey nuances in Japanese, I can't in English. This got me to wondering what, exactly, "nuance" was. It's slightly different from my current research field, but, as a researcher, I am very interested in that which cannot be expressed through written language.

Bono I've always liked speaking publicly, to the degree that I was a member of the drama club while in high school (laugh). In university,I began to wonder "what, really, are

the language expressions that underpin drama", and began studying linguistics. When we talk face-to-face, we use not only audio, but eye contact, body movements, gestures, and the like. In graduate school, my interest began to turn to how we convey thoughts to others through sound and movement.

- Can one clearly separate "language" from body movements and similar "non-language"?

Bono I, myself, was interested in that borderline, pursuing it in my research, but now I have reservations about structurally dividing up all elements. I think that it's necessary to take a comprehensive view that includes both speech and body movement.

Abekawa My approach is a little different from yours, in that I've always liked to read, and always had an interest in the written language. That's what got me into natural language processing research. The first thing I felt acutely was how difficult it is to have computers understand language. For example, in the sentence "I saw someone swimming with grace", grammatically, "grace" modifies "swimming". However, in the sentence "I saw someone swimming with binoculars", "binoculars" does not modify "swimming". Both sentences use the word "with", but they correspond to different words. This difference is clear for humans, but computers, until recently, were unable to comprehend this difference. In the last 10 years or so, the idea has been to teach computers this knowledge and common sense. Perhaps one could compare it with exposing a child to a large number of experiences in order that the child may grow.

Bono Speaking of children, there's something I'd like to mention regarding language acquisition and the environment in which children grow up. For example, when a deaf couple has a hearing child, the child sometimes learns sign language as their first language. The child then goes on to acquire Japanese language ability as a second language, when they start nursery school, or by communicating with people other than their parents. As this shows, language acquisition is heavily dependent on the environment in which a child is raised.

The deaf parents wish the child to understand sign language as part of their identity, but the hearing child comes into contact with Japanese in the form of audio every day. I'm sure there are more than a few families unsure of how to handle this situation.

Facing the World as a Language Professional

Abekawa Professor Bono mentioned "hearing children". I was once involved in Japanese adnominal research, so I unconsciously analyze "hearing" as modifying "child". I often think "that's an odd expression" when I read things others have written.

Uchiyama That's true for me as well. The specialized terminology which I am researching contains many compound words containing multiple words, so when I watch television, my attention is always drawn by "jigyoshiwakesagyo (operation screening work)". Both "screening" and "work" are nouns which come from verbs, and I automatically start to think about the rules behind compound noun creation, and prescribed word order.

Bono What I've been interested in lately is the size of mobile phone and computer screens. Computer screens are large, making it possible to polish ones text, but the screens on mobile phones are small, and polishing ones text is difficult. The stance towards written language probably differs for young people who are used to communicating via mobile phone email and older people who are not.

Abekawa There are even some students who send their graduation thesis to their professors via mobile phone email.

Uchiyama Bono That's unbelievable!

- The three of you, as researchers, stand face to face with the Japanese language and modern society. In closing, could you please discuss how your research can contribute to society?

Abekawa I have placed the focus of my research on how the research results can be applied to actual society. I hope to develop systems which are widely used.

Bono Japanese know greetings in other languages, such as "hello" or "ni hao", but most people don't know how to say "hello" in Japanese sign language. I think that my role is, through the database of Japanese sign language I mentioned earlier, to convey to people another of Japan's languages, Japanese sign language, and get them thinking about "what exactly is language?"

Uchiyama In my interactions with students, I have found that they are interested in search engines, but not in the natural language processing technology used in search engines. I hope to more clearly explain natural language processing, which serves a useful but hidden role in people's lives.

(Written by Yoshihiro Masukuni)

*Associative Search: Search method which selects words which are highly correlated with search keywords, in the same way as people unconsciously think of related words when they hear a word, and use it in searching for books containing said words, etc.

Gossip and Human Bands

Tetsuro Kobayashi
Assistant Professor , Information and Society Research Division, NII

Why makes gossip so fun? There is a significant amount of research that indicates that a large portion of our daily conversations consists of gossiping. There must be some reason that people enjoy gossip so much.

Is Language an Alternative to Grooming?

Anthropologist Robin Dunbar has put forth a bold hypothesis: gossip is necessary for maintaining human bands, and language itself developed in order to convey gossip (Dunbar, 1996). The primates without language faculties maintain band cohesiveness through long hours of grooming that go far beyond necessity. In addition to its hygiene benefits, grooming also provides the social function of reinforcing information necessary for maintaining band cooperative relationships, such as "who I can trust" and "who hasn't returned favors". With the evolution of the cerebral neo-cortex, the size of bands increased, making it impossible to groom most members due to time constraints. Perhaps, then, language was born as an alternative to grooming. By transmitting information such as "who was slacking off during the hunt" and "who was fair in distributing the catch" in the form of rumors, the same social function provided by grooming could be effectively provided. In other words, it is possible that human language itself evolved in order to make it possible to gossip, a necessity for maintaining large bands.

"Gossip" Links People Together

Thought about this way, it seems there is a reason people enjoy gossip. By evaluating others in the form of gossip, and hearing one's own reputation, people can constantly maintain their cooperative relationships within the band. The reason that hearing gossip about an unknown person is not interesting is because gossip is related to maintaining the band. The word "gossip" is derived from "God Sib", that is, "godfather", which was further extended to mean family ties (Kawakami, 1997). In other words, the word gossip was born from a word indicating relationships of profound trust, like those of family. This would appear to be another point of contact between gossip and band maintenance.

By the way, the world is abuzz about SNS and Twitter. Could it be that a fundamental change in human bands is coming due to the increasing speed and scale of Internet gossip? As strange as it may seem, the ancient "logic of bands" may live on in continuously evolving net services.

This month's cover illustration: This month's theme is "language". In the center is the Tower of Babel, which contains countless different languages. A human and computer, pitted in a battle of wits, face each-other across a Scrabble board, drawing languages from the pile. Who will win?!

Weaving Information into Knowledge
NII

NII Today No.34 July 2010 (This English language edition of NII Today corresponds to No.48 of the Japanese edition)

Published by National Institute of Informatics, Research Organization of Information and Systems http://www.nii.ac.jp/
Address: National Center of Sciences 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430
Chief editor: Yoh'ichi Tohkura Cover illustration: Makoto Komori Photography: Hiroyuki Taniguchi Production: Commercial Design Center Inc.
Contact: Publicity and Dissemination Team, Planning and Promotion Strategy Department TEL:+81-3-4212-2131 FAX:+81-3-4212-2150 e-mail: kouhou@nii.ac.jp