# 2 Research Overview of Knowledge Creating Communication Research Center

**KIDAWARA Yutaka**

NICT Knowledge Creating Communication Research Center (KCCRC) developed intelligent communication technology which consist of Multi-language translation, Information analysis and Knowledge processing in the second medium-term plan. MASTAR project accelerate development of practical multi-language speech translation applications and practical information analysis system. ALAGIN forum provide widely our research results to many companies and research institutes. We describes activities of KCCRC.

## 1 Introduction

In the second medium-term plan, the Knowledge Creating Communication Research Center (KCCRC) was established as one of the centers in the Second Research Sector, and thereafter it was reorganized as one of the research centers of the Keihanna Laboratory. The KCCRC planned research and development in natural communication technology, universal content technology, and universal platform technology. A major theme of the current Universal Communication Research Institute is the technological development of universal communications that allow users to overcome the barriers of language, culture and competence. The above plans set a goal for this development and aim to realize an environment that enables people to freely communicate with anybody, anytime and anyplace, in any language, and by any means.

More specifically, the KCCRC develops various communication technologies including multilingual translation, spoken and non-verbal dialogs, collection of reliable information, and intuitive information presentation, in order to build a communication environment on a ubiquitous information communication platform so that any kind of differences in e.g. language, knowledge, competence and the like can be overcome. The Center overcomes various difficulties in the information network and conducts research and development into respective element technologies in seven fields (communication environment, personal adaptation dialog, non-verbal dialog, multi-language spoken dialog, multilingual machine translation, information credibility analysis and knowledge creation from information, and language grids). It then establishes a knowledge-circulation information communication platform.

With respect to the research and development described above, the KCCRC obtained excellent results and attracted global attention. This paper describes an overview of the research conducted at the KCCRC.

## 2 Research and development of natural communication technologies

For the globalization of communication, it is necessary to develop technology with which anyone can easily access necessary information and conduct smooth communication in any language and with any culture without noticing the system working behind it. For this purpose, the aim was to conduct research and development into basic technologies such as language processing technology, language grid development technology, and non-verbal information analysis/utilization technology.

**Language processing and multilingual translation technology**

In order for anyone to receive and transmit information in a natural manner, research and development was conducted into language-handling technology by making an example base collection of about 10 million sentences and a large language dictionary of about 0.4 million words. Research and development was also carried out into high-performance machine translation technology integrating the method and rules of example-based translation and statistical information technology on the basis of the large multi-language research resources that the Center had developed.

For the establishment of the example base, 28 million sentences were collected, much more than the target, to achieve the world's largest example base. To achieve this scale, we not only realized new information processing technologies, such as automatic techniques for crawling and assignment on the Web, but also released a translation Web site "Minna-no Translation (Translation for Everyone)" that incorporated translation support technology by human. This site realized cooperation with volunteer translators, i.e. a Web2.0-like methodology. The example base has thus been established with an increasing number of users.

Regarding high performance multilingual machine translation technologies, we developed travel-conversation machine translation technology with the world's highest accuracy.

This technology was made open to the public as the travel conversation translation system VoiceTra/TexTra, which has attained unprecedented results throughout the world. The VoiceTra achieved 320,000 downloads and 2.7 million accesses in three months, which indicates that the system had a great impact on general users. In order to realize these high accuracy translation technologies, we developed learning translation technology based on statistics and created new technologies for syntax utilization translation, proper noun translation, and multilingual word division. Another noteworthy result obtained with these speech translation technologies was that our activity was designated as a Social Reduction Acceleration Project of the Cabinet Office, Government of Japan. To accelerate the research and development of multilingual speech translation technologies and to promote their practical applications, we launched the Multilingual Advanced Speech and Text Research Project[*1] as a project involving the entire KCCRC. With this project, we accelerated the integration of technologies developed by each laboratory into a speech translation technology to be implemented into society. This technology was then transferred to many companies. With this achievement, our multilingual speech translation technology was ranked as a "top priority" in the 2011-fiscal year budget request from the science technology-related policy priority judgment system[*2].

International cooperation and the standardization of technologies are necessary to promote multilingual speech translation technology throughout the world. For this purpose, the KCCRC promoted the standardization of network-based speech translation systems at ITU-T SG16, and achieved standardization for the first time in the world in a very short period of time. It also started the U-STAR consortium in order to facilitate international research on speech translation using standardized technologies.

[*1] http://mastar.jp/
[*2] http://www8.cao.go.jp/cstp/budget/yusendo_h23/kekka/09-03inv3.pdf

As a language handling technology, the Speech-based Question Answering System "Ikkyu" was developed. This intellectual natural language processing application can search for answers from 600 million pages on the Web to questions which are given to iPhones by voice, and then extracts and presents the answers. We also developed a method of automatically creating a concept dictionary, which represents a network of semantic relations between words, from texts on the Web. With this method, a concept dictionary covering 2.5 million words, much more than the target of 0.4 million words, was automatically created from the Web archives of the above-mentioned 600 million pages. The concept dictionary contains a large amount of knowledge, in one sense, much more knowledge than individuals can have. It contains not only unexpected useful information but also a variety of commonsense knowledge. This dictionary therefore works as a base for various intellectual natural language processing systems including the Speech-based Question Answering System "Ikkyu". On the other hand, syntactic analysis provides a platform for "high-performance machine translation technology" and "language handling technology." Our syntactic analysis of Chinese language achieved the world's highest performance in the three fiscal years from 2008 to 2010, and the developed syntactic analyzer was released in the ALAGIN forum. Morphological analysis also provides the platform and our morphological analysis of Thai and Chinese languages achieved the world's highest accuracy in fiscal years 2008 and 2009. These analysis systems, in combination with systems performing so-called transliteration, have participated in many international contests and have won many first and other prizes.

**Language grid technology**

In order to establish communication with due consideration of cultural backgrounds and to dramatically improve the accessibility and usability of language resources and language processing functions among different cultures, the KCCRC aims to carry out research and de-velopment into technologies that integrate and systemize existing language resources and language processing functions targeting about 10 languages.

For the technologies that integrate and systemize existing language resources and language processing functions, 92 language grid services in 20 languages were provided to 138 organizations in 18 countries, which exceeded the initial goal of the medium-term plan. The federated operation of the language grid with NECTEC facilitates user participation in Asian countries and the language service acquisition of Asian languages. Also since we received requests for cooperation from the European language resources projects MetaNet and ICT-4Law, and from SILT of the United States, we made the software for the language grid available through open source licensing. We thus globally lead the transformation from language resources to language services. We also started an language grid association to support communications based on cultural backgrounds. Through this user-participatory research and development, we have contributed to intercultural collaborations in the medical and educational fields. The Computer Mediated Multilingual Medical Communication Support System, which is an application of the language grid in the medical field, was introduced to Kyoto City Hospital, Kyoto University Hospital and the University of Tokyo Hospital. The Center for Multicultural Information and Assistance Kyoto (NPO), which supported these medical facilities for this project, received an incentive prize from the Minister of State for Special Missions. We also greatly contributed to research communities. For example we started international conferences and held a language grid study meeting at the annual meeting of the Institute of Electronics, Information and Communication Engineers (IEICE). (Every year, about 15 presentations are made by 10 organizations or more). A cloud service of the multilingual collaboration tool "language grid Toolbox" using the language service is being utilized by 30 organizations to support multicultural coexistence and

international interaction activities. It has also been provided for the support of international students at Tokyo University of Foreign Studies and Kyoto University, and for multilingual support for the overseas sections of the IEICE.

### Dialog System

To achieve an information communication system with which people can transfer information appropriately without feeling stressed, the KCCRC aimed to conduct research and development into dialog systems. The systems include, for example, a speech analysis system for communication with network terminals, a non-verbal expression recognition system using facial expressions, gestures, and hand gestures, and a mechanism of information and inference required for dialog.

A large-scale speech translation demonstration experiment was conducted nationwide in Japan to demonstrate our speech processing technology, and log data from 200,000 dialogs was obtained. The feedback from the log data verified the improvement of translation performance. An automatic learning method based on a speech-recognition acoustic model using reliability scores was also developed and its validity was confirmed by the field data. We also experimentally developed a Korean speech recognition system and speech synthesis system to promote multilingualization. These systems achieved world-class performance.

Our spoken dialog technology achieved natural dialog, which sounds like human dialog, by controlling the dialog in accordance with a statistical dialog control model. We also developed a platform called WFSTDM to accomplish dialog control which combines the statistical model and the rules. Furthermore, the technology and a demonstration system were developed to recommend specific sightseeing spots based on an evaluation of users' preferences about sightseeing spots. In order to take account of not only speeches but also a users' body language, we also developed technology to estimate user attributions, interests

using image processing and detecting users responses to inappropriate answers from the system. By integrating this technology with the speech processing technology, we developed a dialog system that could make high-level dialog and verified the usefulness of non-verbal information.

## 3 Research and development for universal content technology

### Knowledge structuring platform technology

This research and development aimed to develop technology to automatically acquire and store explicit knowledge and for the association of this obtained knowledge data. Explicit knowledge includes not only environmental data and spatiotemporal information but also Web information. For this development we analyzed expert technologies in the extracting and structuring of knowledge information and structured knowledge information, and used a function for integrating multiple knowledge structures.

We considered that the cross-sectional links (knowledge links) of knowledge resources among different fields (disasters, natural environment, climate, health, etc.), different media (text, images, etc.), and different sites should be connected to the knowledge of contexts (earth climate change, health and the like) in various ways. For this purpose we developed a "correlation analysis engine" that dynamically identified the correlations among the fields by scanning them. We then analyzed the correlations of information that covers many fields, media and sites in the context mainly of the earth environment, and conducted evaluation experiments to observe the validity of the analysis. We also developed and tested various prototypes using the knowledge-structuring technology of the correlation analysis engine. A Web content browsing method "link-free Web browsing" that utilizes structured knowledge is a new system with new functions and a new user interface. It extracts and shows the "basis" of the knowledge-

link-based connections between Web content to help users understand their correlations and navigate over the knowledge links. We also developed an authoring tool SAVVY WiKi for users to collaboratively edit the correlation knowledge structures of Web content. With this tool and link-free Web browsing, we have completed an environmental evaluation system for editing and browsing next-generation Web content. The knowledge structuring is not limited to languages. We extended it to phenomena that occur at a specific time and in a specific space (such as natural phenomena, social phenomena and the like), and developed a space-time correlation analysis engine (Moving Phenomenon Engine) to accumulate and control knowledge based on its space-time information.

**Information analysis platform technology**

This research and development aims to build an information analysis system for comprehensive information credibility. The system consists of various information analysis technologies for discovering reliable information from Web content and technology for collecting Web information from the Internet.

We also developed information analysis technology for discovering reliable information. It uses an automatic analysis method that ranks 600 million collected Web pages by link analysis (page ranking), text volume, date and time of collection, and selects 100 million pages suitable for information analysis. It then extracts information senders, opinion sentences, main/conflicting expressions, and appearance information, and shows the summaries of the pages. After the extraction of opinion sentences, a corpus was created by classifying the sentences into 7 kinds of opinions from subjective opinions to objective ones. We also developed an automatic extraction technology that used a machine learning method referring to the corpus as training data, and improved extraction accuracy for a variety of topics. Furthermore, a method was developed to cluster the extracted opinion sentences using main expressions and opposing/conflicting expressions. To propose methods for identifying information senders and checking logical consistency, we conducted research and development into an automatic information-sender analysis method based on machine learning techniques by improving the sender analysis model and the opinion analysis model developed by NICT. For the research and development of technologies analyzing false or unreliable information on the network, we also developed a Web collection system to collect not only Web pages but also blogs and news articles, and to update and check the collected information regularly. A method of analyzing the link structure among Web pages was also developed to establish a spam page detection system. The classification accuracy of the analysis method was evaluated for various types of collected documents to improve the accuracy of each automatic function. We incorporated the developed analysis method into the information analysis system WISDOM and improved performance to a practical level in order to give analysis results for unlimited analysis targets. To share this achievement with society, we further improved the system into a form that general users can use, and released it to the public.

**Knowledge cluster platform technology**

This research and development aims at developing a user-oriented information utilization system on international distributed-information analysis architecture. The system collects knowledge that users require from knowledge distributed over many sites. This system was developed as a grid architecture, and a knowledge GRID base has been established for knowledge processing. The base consists of knowledge grid nodes in Japan and overseas (Koganei, Keihanna, Keio SFC, Indonesia, Finland (× 2), South Korea, Germany, China, and other domestic bases (Sapporo, Sendai, Okinawa, etc.). Also, virtual knowledge grid bases were developed and installed in the APs of JGN2plus cities, i.e. Sapporo, Sendai, Tokyo, Keihanna, and Okinawa, and thereby virtual cloud imaging

was deployed. Knowledge services for about 400 fields were built on the knowledge grid bases, and the correlation analysis/search engines that we developed as core technologies were deployed. In these environments, we performed evaluation experiments for cross-field type analysis. We also began research and development into a service search engine, and developed a new search method that took account of service utilization contexts (state of usage of services in various applications), and a prototype of a decentralized distributed indexing/search mechanism.

## 4 Reserch and development of universal platform technology

### User adaptation technology

An objective of this research and development was to develop a system that provided suitable information to each user by enhancing the robustness of the real-time sensing technology that detects a user's non-verbal information (face direction, visual line, facial expressions, body action, etc.) against environmental fluctuations. For the research and development of real-time sensing technology to detect the non-verbal information of users, which was a core technology of this system, we relaxed the restrictions on illumination and integrated the information data obtained from multiple sensors (three cameras) to acquire non-verbal information with a high degree of accuracy even without using special illumination. We also developed a system that accurately detected the positions of multiple persons in front of the system using a color stereo camera and estimated the head position of each person. We used these technologies to develop a prototype of a large screen dialog system, where not only voice information but also non-verbal information such as the identification of people and face direction could be acquired. This system can control dialog based on user face direction. We conducted demonstration tests for a total of 100 persons and created an evaluation database of 12,000 dialogs of 20 hours.

### Regional adaptive communication platform

In order to keep an eye on or care for, in particular, elderly people at home, this research and development aims to develop a home sensor network technology for monitoring the living status of residents, and flexible technology for sending and receiving information in accordance with the monitored status. This research and development has been transferred to the Medical ICT Group, New Generation Wireless Communications Research Center, during the term of the medium-term plan.

In this research and development into regional adaptive communication platform technologies, we promoted the development of high-speed communication technology and applied technology for two-dimensional communications. We also developed a system to receive carrier signals (pilot signals) from multiple couplers generated at terminals placed on a two-dimensional communication sheet, to measure and compare the intensities and phases of the signals, and to automatically concentrate the power to a client terminal placed at an arbitrary location. We also proposed a system that provides a time-shared power supply to multiple terminals according to the combination of multiple phase assignment settings, and the one that provides simultaneous power supply to multiple terminals according to a single phase assignment setting.

## 5 Contribution to society via ALAGIN Forum

The Advanced Language Information Forum (called "ALAGIN forum")[*3] was established in 2008 with the aim of collaborations between government, industry, and academia for speech language processing technologies, and has achieved many enrollments from companies and universities. The number of end-user license agreements for the linguistic resources and services that NICT releases at

---

＊3 http://www.alagin.jp

ALAGIN is now more than 380 as of the end of the 2009 fiscal year (last year of the second medium-term plan). In the ALAGIN Forum, not only technologies and data related to the above-mentioned concept dictionary but also a large amount of data and software such as the bilingual corpus, Japanese language WordNet, and Chinese language analysis tool are released. Some data has been downloaded more than 8,000 times. In some cases, the downloaded data is now being used in iPhone applications and Web services.

## 6 Conclusions

The research and development of the Knowledge Creating Communication Research Center is at the world's highest level. It has developed technologies such as practical multi-language translation and information analysis, which had not existed before.

The KCCRC also provided large-scale language resources highly useful for companies and research institutes and created technologies useful for society. It has also made many presentations at high-level international conferences. The number of presentations that the KCCRC has made is the highest among research institutes in Japan. In particular, at a great international conference in 2010 for the language processing of ACL and EMNP, we presented the 4th most papers in the world's. The KCCRC has thus created results at quite a high level even from an academic viewpoint. For speech translation, we have jointly held the international workshop IWSLT every year since 2004, leading worldwide speech translation studies. Through such research and development, we have been awarded many academic prizes such as the Japan Society Award for the Promotion of Science, the 43rd Ichimura Science Award, and the 56th Maejima Award.

Universal communication technology to overcome barriers of communication, which is an objective of the research and development at the KCCRC, became the name of the research center in the third medium-term plan, and its research and development was acceler-

ated. The translation technology for which research and development was performed by the Language Translation Group is now studied in the Multilingual Translation Laboratory for worldwide applications. The speech processing technology is indispensable for not only future information access but also the multimedia data processing of video images recorded on YouTube and the like. Therefore the Spoken Language Communication Laboratory has conducted research and development into speech recognition, speech dialog, and speech synthesis in order to accelerate studies on speech processing technologies. The Language Infrastructure Group and the Knowledge Clustered Group have studied information analysis technologies and information utilization technologies. However in the third medium-term plan, these groups were reorganized into the Information Analysis Laboratory and Information Services Platform Laboratory, respectively. Using previously developed technologies as core technologies, these Laboratories are conducting research and development of information services platform technologies, which are expected to play a role as a service base for larger scale language processing and for the cross-sectional processing of information in different fields. New research and development fields of the NICT created in the second medium-term plan are generating more great results in the third medium-term plan. The NICT has become a laboratory that attracts a lot of attention from around the world.

**KIDAWARA Yutaka**, *Ph.D.*

*Director General, Universal Communication Research Institute*

*Digital Content Management, Ubiquitous Computing, Information Retrieval, Information Analysis*