Media Processing Technology for Achieving Hospitality while on the Go

Motoyuki Horii, Kazuhiro Arai, Masaaki Nagata, Kunio Kashino, Kaoru Hiramatsu, Atsushi Fukayama, and Hitoshi Yamaguchi

Abstract

This article introduces a service concept of *hospitality on the go*, in which users are guided while they move around a local area, as well as statistical machine translation technology and robust media search technology for supporting such services.

Keywords: statistical machine translation, robust media search, hospitality

1. Introduction

With the year 2020 in mind, NTT's goal is to implement a navigation service that can provide foreign visitors to Japan, who are moving about a local area, with detailed guidance according to the user's attributes and situation. This concept is described in detail below.

1.1 Service for guiding people to their destination in unfamiliar places

In recent years, the translation of Japanese guidance information into other languages on information displays in public transportation facilities such as train stations has been progressing. Nevertheless, information that can change at any time due to delays or accidents cannot be translated in advance. Moreover, visitors from other countries do not have an intuitive understanding of the local geography, so simply translating the names of places and exits that appear on signs does not help them decide which way to go.

To achieve a more effective navigation aid, detailed information presented in Japanese is translated in real time by using multilingual statistical translation technology, and appropriate guidance information is selected based on an estimation of the user's situation (**Fig. 1**). Robust media search (RMS) technology, which can recognize objects that the user sees around them, and various other types of recognition technology are used to estimate the user's situation. For example, when a train station employee inputs emergency information in Japanese, that information is immediately translated and sent to the smartphones of foreign visitors who are in that station, and digital signage or message boards can be translated and displayed on smartphones held up to the boards by users. It would also be possible to display navigation instructions to destinations in various languages by interworking with smartphones.

1.2 Tourist navigation service based on "What I can see now"

At tourist sites and places being visited for the first time, the surroundings as seen by the user are captured with a smartphone or head-mounted device, and that information is used to provide location-based guidance. Video of scenery can include various photographic angles and objects in the environment. RMS-object is a specialized type of object recognition technology for recognizing photographic subjects. RMS-object can be used to discover multiple objects from different viewing angles and environments with higher accuracy (**Fig. 2**). Combining RMS-object with technology capable of estimating the user's situation makes it possible to provide

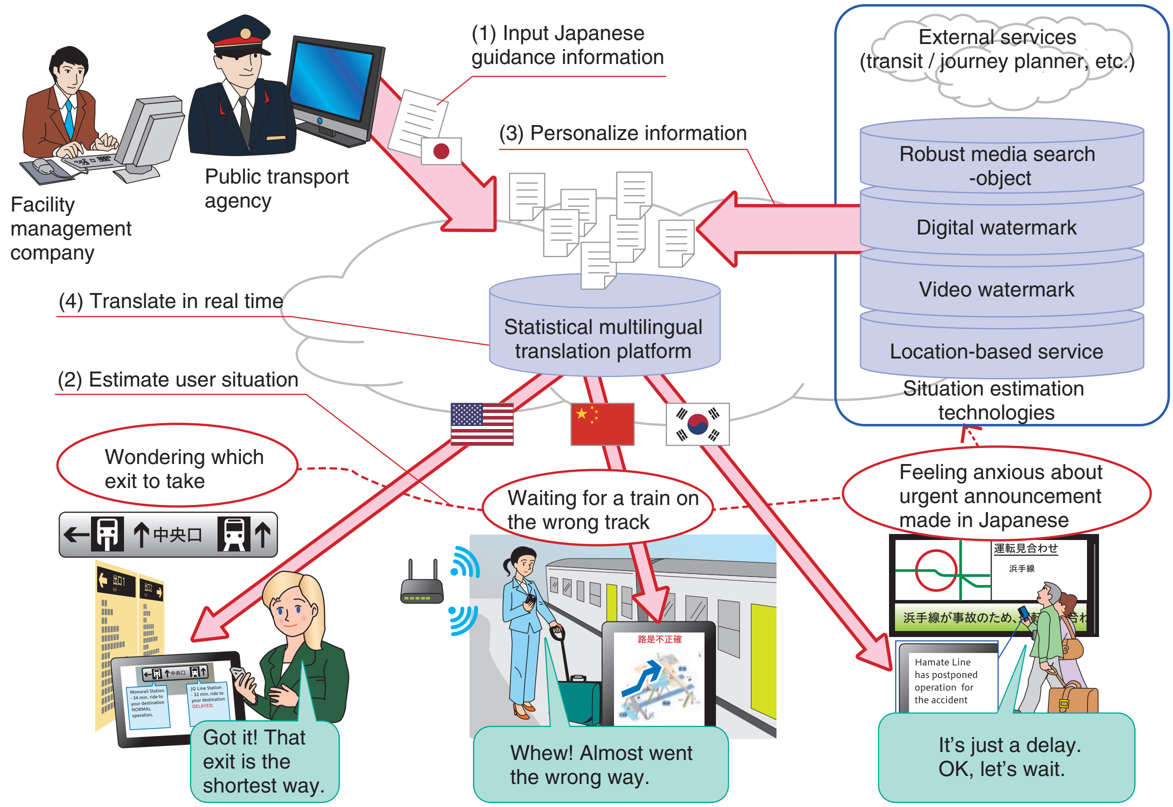

Facility management company
Public transport agency
(1) Input Japanese guidance information
(3) Personalize information
External services (transit / journey planner, etc.)
Robust media search -object
Digital watermark
Video watermark
Location-based service
Situation estimation technologies
(4) Translate in real time
Statistical multilingual translation platform
(2) Estimate user situation
Wondering which exit to take
Waiting for a train on the wrong track
Feeling anxious about urgent announcement made in Japanese
←中央口↑
出口1
出口2
Monorail Station - 34 min. ride to your destination NORMAL operation.
JQ Line Station - 32 min. ride to your destination DELAYED.
路是不正確
運転見合わせ
浜手線
浜手線が事故のため、
Hamate Line has postponed operation for the accident
Got it! That exit is the shortest way.
Whew! Almost went the wrong way.
It’s just a delay. OK, let’s wait.

Fig. 1. Hospitable navigation (public transportation).

guidance according to the user’s location at the time by selecting and displaying what is appropriate for the user’s attributes and situation from the information available on the discovered objects.

2. Media processing technology to support *hospitality on the go*

NTT is moving forward with research and development (R&D) of statistical machine translation and RMS technology to be applied in implementing hospitality on the go services.

2.1 Statistical machine translation

As use of the Internet increases and the reach of globalization spreads further, the need for language translation done by computers, known as machine translation, is also increasing. Work on machine translation to eliminate the barrier of language, including work on a national level, is accelerating as we look toward 2020, and expectations are high.

R&D on machine translation has a long history, and many machine translation systems have already been developed. Nevertheless, existing systems have not really met worldwide needs and expectations, so an innovative advance in technology is needed.

Conventional machine translation systems that use a rule-based translation approach require years of work by many experts to manually produce translation rules and bilingual dictionaries for translation of a new language. Such systems have already reached the limit of accuracy of manual work, and thus, in recent years, a different approach called statistical machine translation has become mainstream. In this approach, a statistical model that is equivalent to translation rules and bilingual dictionaries is learned automatically from large-scale bilingual data on the order of several million sentence pairs.

The outline of statistical machine translation is illustrated in **Fig. 3**. It achieved at an early stage a practical level of accuracy for language pairs that have very similar word orders such as English and French. For languages that have greatly different word orders such as English and Japanese, however, it was not able to outperform the conventional rule-based translation.

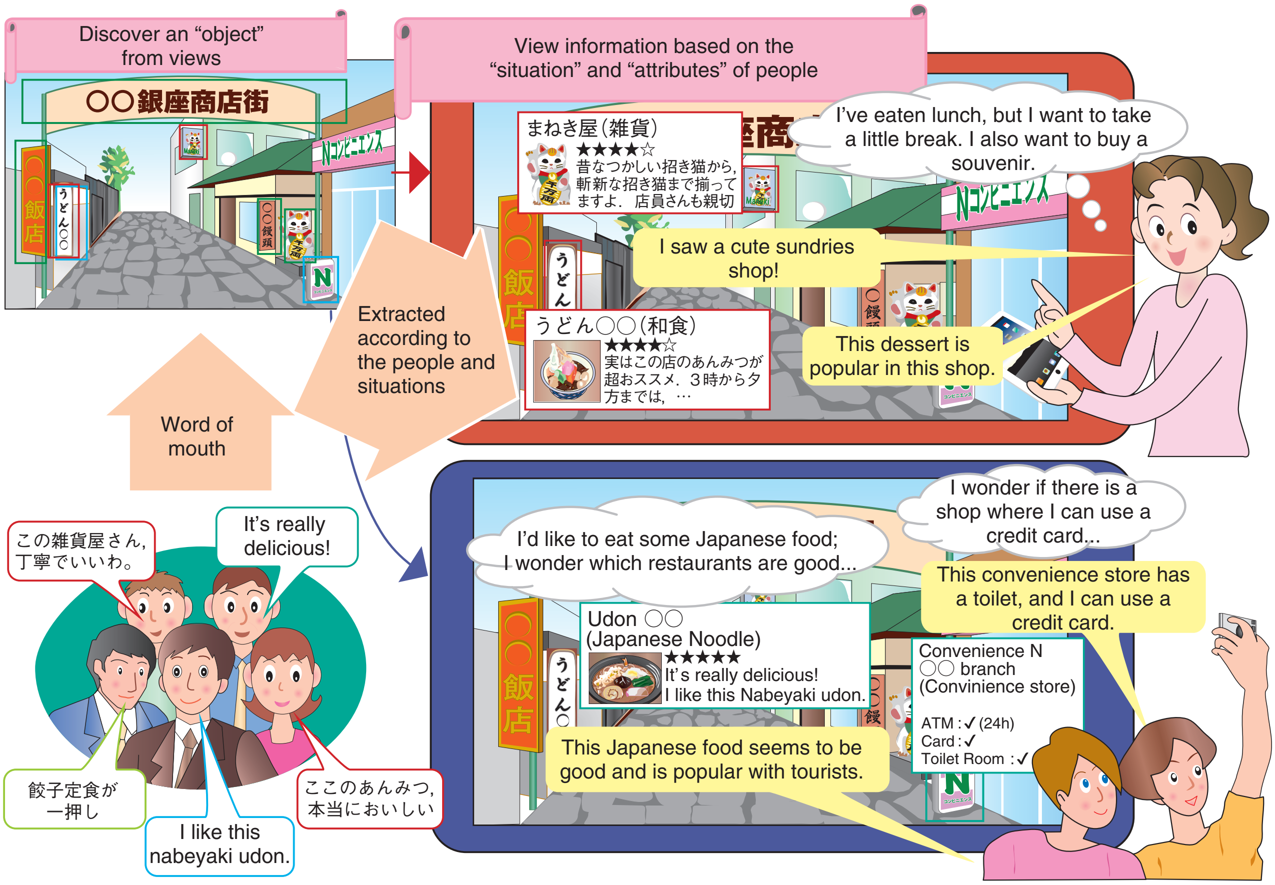

Discover an “object” from views
〇〇銀座商店街
Nコンビニエンス
〇〇飯店
うどん〇〇
View information based on the “situation” and “attributes” of people
まねき屋（雑貨）
★★★★☆
昔なつかしい招き猫から，斬新な招き猫まで揃ってますよ．店員さんも親切
I’ve eaten lunch, but I want to take a little break. I also want to buy a souvenir.
I saw a cute sundries shop!
うどん〇〇（和食）
★★★★☆
実はこの店のあんみつが超おススメ．３時から夕方までは，…
This dessert is popular in this shop.
Extracted according to the people and situations
Word of mouth
この雑貨屋さん，丁寧でいいわ。
It’s really delicious!
餃子定食が一押し
I like this nabeyaki udon.
ここのあんみつ，本当においしい
I’d like to eat some Japanese food; I wonder which restaurants are good...
I wonder if there is a shop where I can use a credit card...
This convenience store has a toilet, and I can use a credit card.
Udon 〇〇 (Japanese Noodle)
★★★★★
It’s really delicious!
I like this Nabeyaki udon.
Convenience N 〇〇 branch (Convinience store)
ATM : ✓ (24h)
Card : ✓
Toilet Room : ✓
This Japanese food seems to be good and is popular with tourists.

Fig. 2. Hospitable navigation (tourist site).

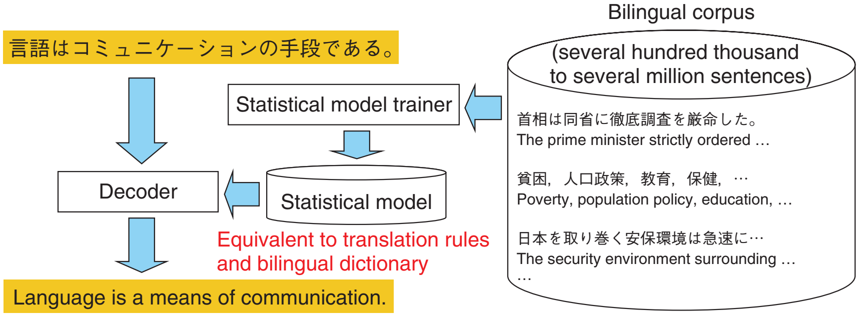
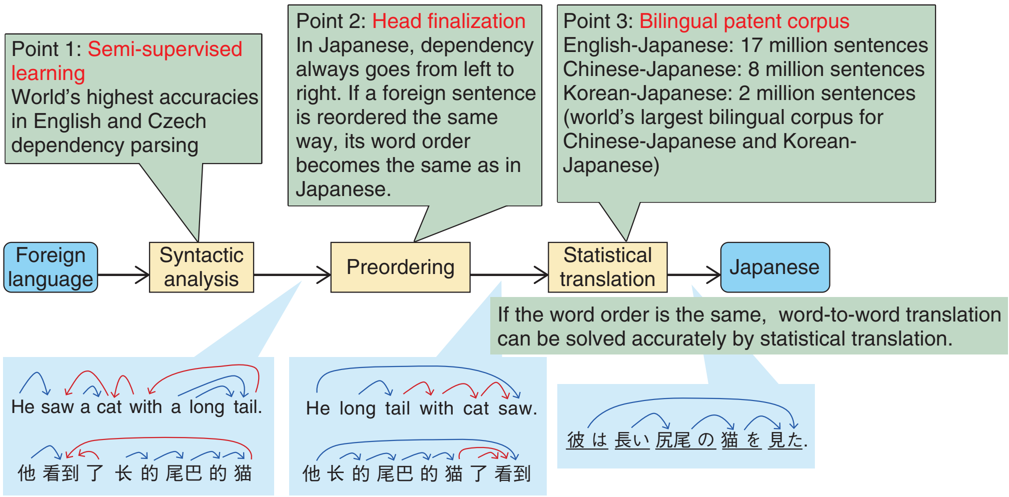

言語はコミュニケーションの手段である。
Statistical model trainer
Decoder
Statistical model
Equivalent to translation rules and bilingual dictionary
Language is a means of communication.
Bilingual corpus
(several hundred thousand to several million sentences)
首相は同省に徹底調査を厳命した。
The prime minister strictly ordered …
貧困，人口政策，教育，保健，…
Poverty, population policy, education, …
日本を取り巻く安保環境は急速に…
The security environment surrounding …
…

Fig. 3. Outline of statistical machine translation.

NTT has devised a method in which statistical machine translation is applied after reordering English words into Japanese word order. The reordering of English words is based on the single Japanese linguistic property of head finality [1]. For the first time in English-to-Japanese translation, we achieved a result in which statistical machine translation outperformed rule-based translation in accuracy [2].

The concept of word reordering based on the Japanese head-final property is illustrated in **Fig. 4**. The word that determines the grammatical role of a phrase in a sentence is called the *head*; or, as is learned in



Point 1: Semi-supervised learning
World's highest accuracies in English and Czech dependency parsing

Point 2: Head finalization
In Japanese, dependency always goes from left to right. If a foreign sentence is reordered the same way, its word order becomes the same as in Japanese.

Point 3: Bilingual patent corpus
English-Japanese: 17 million sentences
Chinese-Japanese: 8 million sentences
Korean-Japanese: 2 million sentences
(world's largest bilingual corpus for Chinese-Japanese and Korean-Japanese)

If the word order is the same, word-to-word translation can be solved accurately by statistical translation.

Foreign language
Syntactic analysis
Preordering
Statistical translation
Japanese
He saw a cat with a long tail.
他 看到 了 长 的 尾巴 的 猫
He long tail with cat saw.
他 长 的 尾巴 的 猫 了 看到
彼 は 長い 尻尾 の 猫 を 見た.

Fig. 4. Preordering based on Japanese head-final property.

Japanese elementary school, the word that is modified by others is the head. In Japanese, the dependency always goes from left to right, which is to say that the modified word is always placed at the sentence-end side. The term that describes this relationship is called *head finality*. Therefore, if we reorder the words of the translation source language (English or Chinese) so that its dependency always goes from left to right, the resulting word order becomes the same as the Japanese word order. If the word order is the same, highly accurate translation is possible through literal word-by-word translation.

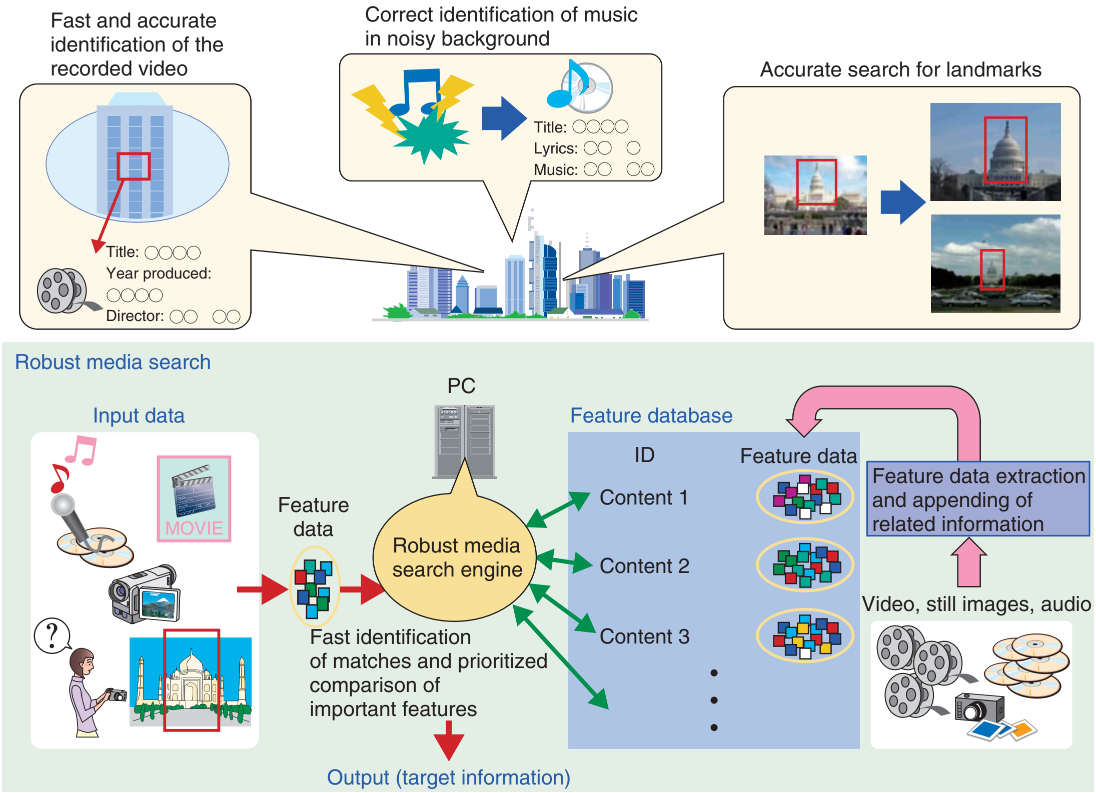
Translating from Japanese to foreign languages (English or Chinese), on the other hand, is difficult because we must select the dependency relations in the input Japanese sentence that should be reversed from right to left based on the word order of the target language.

NTT has devised a translation method for changing the word order of Japanese sentences to that of the target language based on the predicate-argument structure of Japanese [3]. The predicate-argument structure is the grammatical relationship between a verb and nouns, namely, which noun is the subject of the verb and which noun is the object of the verb. As is learned in English classes in Japanese middle schools, English has an SVO (subject, verb, object) word order, while Japanese has an SOV (subject, object, verb) order.

Therefore, we first identify the predicate-argument structure of the Japanese sentence and then change the order of the *bunsetsu* phrases (roughly corresponding to noun phrases, prepositional phrases, verb phrases, etc. in English) to convert the Japanese SOV order to the SVO order of English. Because English and Japanese have an opposite word order within a bunsetsu phrase, the next step is to reorder the words in the Japanese bunsetsu phrases to match the English order (e.g., 東京で→in Tokyo). This approach reduces the number of word ordering errors in statistical translation from Japanese to English by about 30% relative to the conventional method.

The Multilingual Statistical Translation Platform (PF) was developed with the machine translation technology described above. The platform currently handles translation from English, Chinese, and Korean to Japanese and from Japanese to those languages. In addition to the main translation function, the platform provides a user dictionary function, an unknown language detection function, and other functions that are needed for business applications. Functions for user convenience, such as support for creating statistical models, which is difficult for ordinary users, are also implemented.

The quality of statistical machine translation depends on the amount of data used to train the statistical model. We achieved high-quality in translating patents by using this platform with large datasets of corresponding sentence pairs that we created from patent documents for English and Japanese (about 17 million sentence pairs), Chinese and Japanese (about 8 million sentence pairs), and Korean and Japanese (about 2 million sentence pairs). Replacing the data used when training the statistical model makes it


Fast and accurate identification of the recorded video
Title: ○○○○
Year produced: ○○○○
Director: ○○ ○○
Correct identification of music in noisy background
Title: ○○○○
Lyrics: ○○ ○
Music: ○○ ○○
Accurate search for landmarks
Robust media search
Input data
MOVIE
PC
Feature data
Robust media search engine
Fast identification of matches and prioritized comparison of important features
Output (target information)
Feature database
ID
Feature data
Content 1
Content 2
Content 3
Feature data extraction and appending of related information
Video, still images, audio

ID: identification
PC: personal computer

Fig. 5. Robust media search technology.

possible to automatically construct high-quality machine translation systems for particular domains other than patents.

The development of innovative machine translation technology and its implementation in a system as described above has laid the foundation for meeting worldwide needs and expectations. In the future, we will aim for even higher accuracy and a wider range of target domains to achieve machine translation that truly removes the barrier of language.

2.2 RMS

RMS is technology for using small parts of video or still-image signals from a camera or sound signals from a microphone as keys to search a large database that contains videos, music, and still images of landmarks (**Fig. 5**) [4, 5].

This kind of matching-based media search has been the subject of R&D by NTT laboratories for over 20 years, and the results have served as the core of various services, including *net monitoring* for investigating the use of video on the Internet, *music use listing* for automatically creating lists of music used in broadcast programming, and *second screen* for displaying network content related to broadcast programming on smartphones by capturing audio or video in the programming.

RMS is robust against ambient noise or obstacles, distortion in video, and interruptions in audio, and it can also search huge amounts of media data instantly. For example, it is possible to identify the name of a song that can be heard amidst street noise. For video, it is possible to quickly and accurately identify objects that are partially hidden and cannot be seen in their entirety. Because RMS uses video and sound rather than textual information, it is possible to identify what is seen or heard when the names of those things are not known by the user or when text input is difficult. For hospitality while on the go, this function can be used to recognize objects that can be seen in the surroundings and to display appropriate information according to the user’s attributes and situation.

We are currently working on increasing the speed and accuracy as well as the ease of use of RMS to enable fast and accurate searches at the moment and on the spot. In the future, we will continue to study actual use environments and to do basic research on media search technology.

3. Future development

The idea of hospitality on the go places importance on understanding the user's situation and intentions, which is one of the main elements of our vision of a personal agent. People who visit Japan find it difficult to gather information in public places because of language differences and indecipherable signs and information displays. Our approach to overcoming this problem is to provide that information in different forms that suit users and to display it in useful ways.

Our goal for the future is to implement services to provide a better user experience by going beyond the translation and video search technology described in this article through interworking with technology related to geographic data and other important technical elements.

References

[1] H. Isozaki, K. Sudoh, H. Tsukada and K. Duh, "HPSG-based Preprocessing for English-to-Japanese Translation," ACM TALIP. Vol. 11, No. 3, Sept. 2012.

[2] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou, "Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop," Proc. of NTCIR-9, pp. 559–578, Tokyo, Japan, Dec. 2011.

[3] S. Hoshino, Y. Miyao, K. Sudoh, and M. Nagata, "Two-state Preordering for Japanese-to-English Statistical Machine Translation," Proc. of IJCNLP 2013, pp. 1062–1066, Nagoya, Japan, Oct. 2013.

[4] K. Kashino, R. Mukai, K. Otsuka, H. Nagano, T. Izumitani, A. Kimura, T. Kurozumi, and J. Yamato, "Fast Media Search," NTT Technical Journal, Vol. 19, No. 6, pp. 29–32, 2007 (in Japanese).

[5] K. Kashino, "Search and Use of Huge Amounts of Media Data—towards Eliminating the Bottleneck of the Big Media Era," NTT Technical Journal, Vol. 26, No. 4, pp. 31–34, 2014 (in Japanese).

Motoyuki Horii

Senior Research Engineer, Supervisor, Project Leader of Natural Language Processing Systems Development Project, Promotion Project 1, NTT Media Intelligence Laboratories.

He received the B.S. and M.S. in computer science from Keio University, Kanagawa, in 1986 and 1988, respectively. Since joining NTT in 1988, he has mainly been engaged in R&D of natural language processing, agent communication, IPTV systems, and multimedia broadcasting systems. He is currently developing multilingual statistical machine translation systems and introducing them to NTT Group companies.

Kazuhiro Arai

Senior Research Engineer, Supervisor, Promotion Project 1, NTT Media Intelligence Laboratories.

He received the B.E. and M.E. in electrical engineering from Kansai University, Osaka, in 1987 and 1989, respectively, and the Ph.D. in information engineering from Osaka University in 1992. He joined NTT Human Interface Laboratories in 1992 and studied dialogue controls on spoken dialogue systems. During 1996–1997, he was a visiting researcher at AT&T Laboratories in Florham Park, NJ, USA. In 1999, He moved to NTT Communications and joined a commercial system development project. He developed the Voice Portal service system and softphone for voice-over-IP services. During 2009–2014, he promoted the Voice Mining platform in NTT Media Intelligence Laboratories. He is currently promoting the Multilingual Machine Translation platform.

Masaaki Nagata

Senior Distinguished Researcher, Group Leader, NTT Communication Science Laboratories.

He received the B.E., M.E., and Ph.D. in information science from Kyoto University in 1985, 1987, and 1999. He joined NTT in 1987. He was with Advanced Telecommunications Research Institute International (ATR), Interpreting Telephony Research Laboratories, Kyoto, from 1989 to 1993. He was a visiting researcher at AT&T Laboratories, NJ, USA, from 1999 to 2000. His research interests include natural language processing, especially morphological analysis, named entity recognition, parsing, and machine translation. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Information Processing Society of Japan (IPSJ), the Japanese Society for Artificial Intelligence (JSAI), the Association for Natural Language Processing (ANLP), and the Association for Computational Linguistics (ACL).

Kunio Kashino

Senior Research Scientist, Supervisor, Distinguished Researcher, and Leader of Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the Ph.D. from the University of Tokyo for his pioneering work on music scene analysis in 1995. Since joining NTT in 1995, he has been working on audio and video analysis, search, retrieval, and recognition algorithms and their implementation. He has received several awards including the Maejima Award in 2010, the Young Scientists' Prize for Commendation for Science and Technology from the Ministry of Education, Culture, Sports, Science and Technology in 2007, and the IEEE (Institute of Electrical and Electronics Engineers) Transactions on Multimedia Paper Award in 2004. He is a senior member of IEEE. He is also a Visiting Professor at the National Institute of Informatics, Tokyo.

Kaoru Hiramatsu

Senior Research Scientist, Supervisor, Media Recognition Research Group NTT Communication Science Laboratories.

He received the B.S. in electrical engineering and the M.S. in computer science from Keio University, Kanagawa, in 1994 and 1996, respectively, and the Ph.D. in informatics from Kyoto University in 2002. In 1996, he joined NTT Communication Science Laboratories and has been working on the Semantic Web, sensor networks, and media search technology. From 2003 to 2004, he was a visiting research scientist at the Maryland Information and Network Dynamics Laboratory, University of Maryland, USA. He is a member of IPSJ and JSAI.

Atsushi Fukayama

Senior Research Engineer, 2020 Epoch-making Project, NTT Service Evolution Laboratories.

He received the B.S. and M.S. in precision engineering from Kyoto University in 1997 and 1999. He joined NTT laboratories in 1999 and has been engaged in basic research on image recognition, human computer interaction, and application service development utilizing technologies such as network storage and augmented reality. He is a member of IEICE.

Hitoshi Yamaguchi

Senior Research Engineer, 2020 Epoch-making Project, NTT Service Evolution Laboratories.

He received the B.S. in physics from Keio University, Kanagawa, in 1997 and the M.S. in applied physics from Tokyo Institute of Technology in 1999. He joined NTT Network Service Systems Laboratories in 1999 and has been engaged in research on network and communication services. He is a member of IEICE and JSAI.