A Cross-lingual Communication System with Agent-mediated Architecture

Naoki Asanoma†, Setsuo Yamada, Osamu Furuse, Masahiro Oku, Akira Kataoka, and Yamato Takahashi

Abstract

This paper describes a speech translation system with a new architecture for goal-oriented communication. In this architecture, an agent that understands the goal of communication mediates between two participants who speak different languages. The agent guides the participant's utterances to precisely acquire important keywords in goal-oriented communication. The agent then generates the corresponding utterance in the other language by embedding the translated keywords into a sentence template that has been prepared in advance. This approach enables us to achieve the goal of communication using current speech recognition and machine translation technologies, even though their accuracy is not yet considered sufficient for practical use.

1. Introduction

Several speech translation systems have been developed [1]-[5] that attempt to enable natural communication between people who speak different languages. These systems concatenate speech recognition (SR) and machine translation (MT) modules. Many attempts have been made to improve the accuracies of both kinds of module. For example, Sumita [6] proposed a technique to improve the MT module by using large-scale bilingual corpora and multiple MT engines. However, current SR and MT technologies do not provide sufficient accuracy for practical use. It is difficult for speech translation systems to recognize and translate correctly the various kinds of expressions common in spontaneous utterances.

Some attempts have been made to handle spontaneous utterances in speech translation systems by creating systems that interact with the users [7]-[9]. In these attempts, the candidate sentences of SR results are shown to the user before translation and/or the results of the syntactic or semantic analysis performed as part of the MT process are shown to the user for confirmation. Another system uses a dialog management mechanism to guide the interactions to achieve correct translations [5]. Unfortunately, they do not completely eliminate miscommunication.

To avoid miscommunication, we propose an agent-mediated architecture for speech translation systems to achieve goal-oriented communication. This agent guides participants toward their goals through interactions that focus on important keywords. The architecture differs from other speech translation systems because our system requires only enough interaction to acquire the information needed to achieve the goal.

2. Spoken translation system with agent-mediated architecture

2.1 System framework

Figure 1 shows a block diagram of our system based on the agent-mediated architecture for goal-oriented cross-lingual communication. The two participants do not communicate directly with each other. Utterances are neither directly translated nor conveyed to the other side, unlike in other speech

† NTT Cyber Solutions Laboratories
Yokosuka-shi, 239-0847 Japan
E-mail: asanoma.naoki@lab.ntt.co.jp

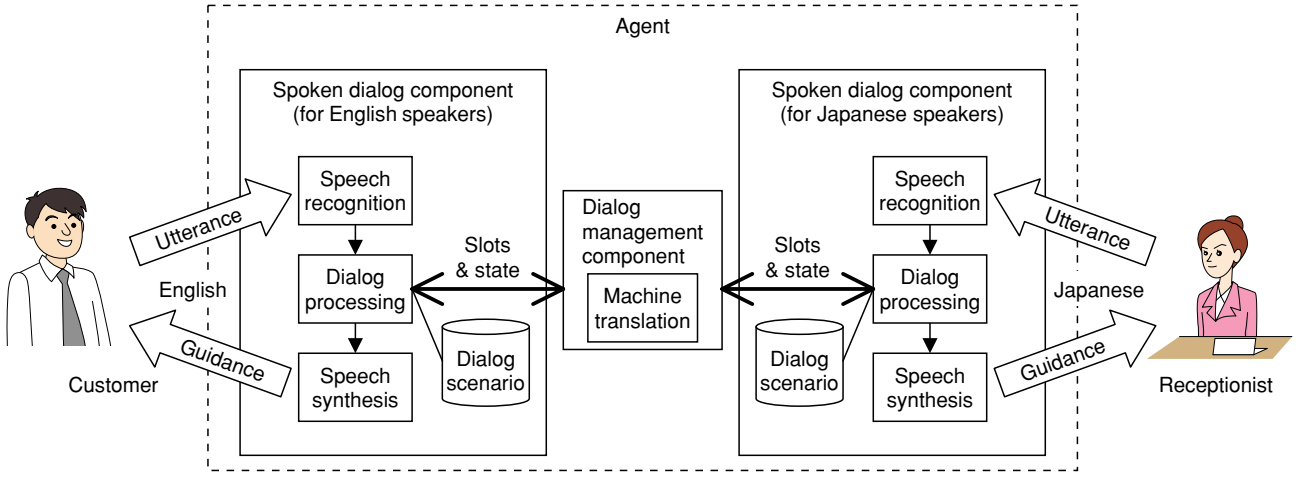

Fig. 1. Block diagram of our architecture.

translation systems. Instead, to acquire important keywords, a mediation mechanism like interpretation (hereafter called the "agent") for their communication is introduced. The agent consists of two spoken-dialog components and a dialog management component. A participant interacts with one of the spoken-dialog components in his/her own language. The interactions follow a prepared dialog scenario that describes how to acquire important keywords in goal-oriented communication.

2.2 Interaction between spoken dialog component and participant

Spoken-dialog component

In our implementation, we used a spoken-dialog component devised for monolingual spoken dialogs [10], such as those used in speech-driven information retrieval services or speech-driven reservation services. It consists of an SR module, a dialog processing (DP) module, and a speech synthesis module, as shown in Fig. 1. The SR module converts the participant's utterance into a word sequence. The DP module extracts one or more keywords from the SR output and then decides the next action according to the dialog scenario. The speech synthesis module provides agent guidance.

Dialog scenario

Table 1 shows a typical dialog scenario that consists of a set of dialog states. In each dialog state, the spoken-dialog component determines what guidance should be given by the agent to the participant and what the next dialog state should be according to the acquired domain slots, which are filled with the keywords extracted by the DP module. For example,

Table 1. Part of the customer-side scenario for making a restaurant reservation.

State	Template sentence of agent guidance	Domain slot	Next state	Submit info.
A10 (initial)	What date and time would you like to reserve?	**[Date]** and **[Time]**	A20	—
		[Date]	A11	—
		[Time]	A12	—
A11	What time would you like on **[Date]** ?	**[Date]**	A11	—
		[Time]	A20	—
⋮	⋮	⋮	⋮	⋮
A20	May I tell the receptionist that your request is for **[Time]** on **[Date]** ?	**[Yes]**	Submit	**[Date]**, **[Time]** B10
		[Time] or **[Date]**	A20	—
⋮	⋮	⋮	⋮	⋮
A30	Your request has been accepted. May I have your name please?	**[Name]**	A31	—
⋮	⋮	⋮	⋮	⋮

when the goal is to make a restaurant reservation, the domain slot for [Date] is filled with a keyword expressing a date, such as "tomorrow" or "May 1st". We determine in advance the types of domain slots, such as [Date], needed to achieve the goal of communication. To precisely acquire the domain slot entries, i.e., keywords, for goal-oriented communication, the dialog scenario is set as follows:

- If a domain slot is not filled, the spoken-dialog component guides the participant to make an utterance that contains the domain slot type needed.
- If the domain slot has been filled, the spoken-dialog component confirms the accuracy of the domain slot entry with the participant.

After all slots have been filled and confirmed, the spoken-dialog component submits the domain slots to the dialog management component with their entries, including the dialog state of the other spoken-dialog component.

Collaboration between two spoken dialog components

The participants' utterances and intentions are exchanged via the dialog management component. The dialog management component controls the two spoken-dialog components and has a bidirectional MT module. After one spoken-dialog component submits the domain slots with their entries and the dialog state of the other spoken-dialog component, the dialog management component translates the keywords using the MT module immediately. The management component submits the translation results to the other spoken-dialog component, which then commences its actions from the submitted state according to the dialog scenario.

Sentence templates are used to convey the submitted keywords to the other participant. The other spoken-dialog component generates agent guidance by embedding the translation results into the appropriate sentence template. The other component starts to interact with the other participant by vocalizing the generated sentence.

One spoken-dialog component waits until the other spoken-dialog component submits the other participant's reply to the dialog management component.

3. Advantages of agent-mediated architecture

The advantages of our architecture can be seen in the dialog example in **Fig. 2**. In this dialog, the English-speaking customer wants to make a reservation at a restaurant whose receptionist speaks only Japanese. To achieve this goal, the customer must let the receptionist know the desired date and time, and the receptionist must confirm whether or not the request has been accepted. The dialog is controlled according to dialog scenarios, parts of which are shown in **Tables 1** and **2**. In these tables, the bracketed words, such as [Time], indicate domain slots. The customer-side scenario (Table 1) regulates what utterance the customer-side spoken dialog component should give the customer in each dialog state and which dialog state should be selected next. If the customer confirms the keywords in the domain slots in dialog state A20, the domain slots with their keywords and the dialog state B10 of the restaurant-side component are submitted to the dialog management component. The restaurant-side scenario (Table 2) regulates the actions of the restaurant-side component.

There are three advantages to our agent-mediated architecture for goal-oriented cross-lingual communication.

(1) Narrowing down the participant's utterances:
It can narrow down the participants' utterances to one of a few known responses by asking carefully designed agent-guided questions. For example, the agent asks the customer a question to acquire the date and/or the time at the beginning of the dialog in Fig. 2. The customer is expected to respond with the date and/or the time of the reservation. This narrows down the number of expected expressions, which improves SR accuracy.

(2) Acquisition of precise information:
It can confirm whether an agent has acquired the information necessary to achieve the goal. As shown in dialog fragment (a) in Fig. 2, despite the agent guidance to acquire both the date and the time, the customer supplies only the date, so the agent queries the time again.

Furthermore, the SR results sometimes include errors due to extraneous utterance parts. However, people tend to ignore such unimportant elements, and instead focus on a few keywords. Since the agent knows which keywords are associated with which goals, it can acquire and confirm all the necessary keywords by repeatedly querying the participant. In the event of an SR error, the customer can check and correct the errors, as shown in dialog fragment (b) in Fig. 2. The customer corrects the time, so the domain slot [Time] is updated with the new time and reconfirmed with the customer. Through these interactions, the agent can precise-

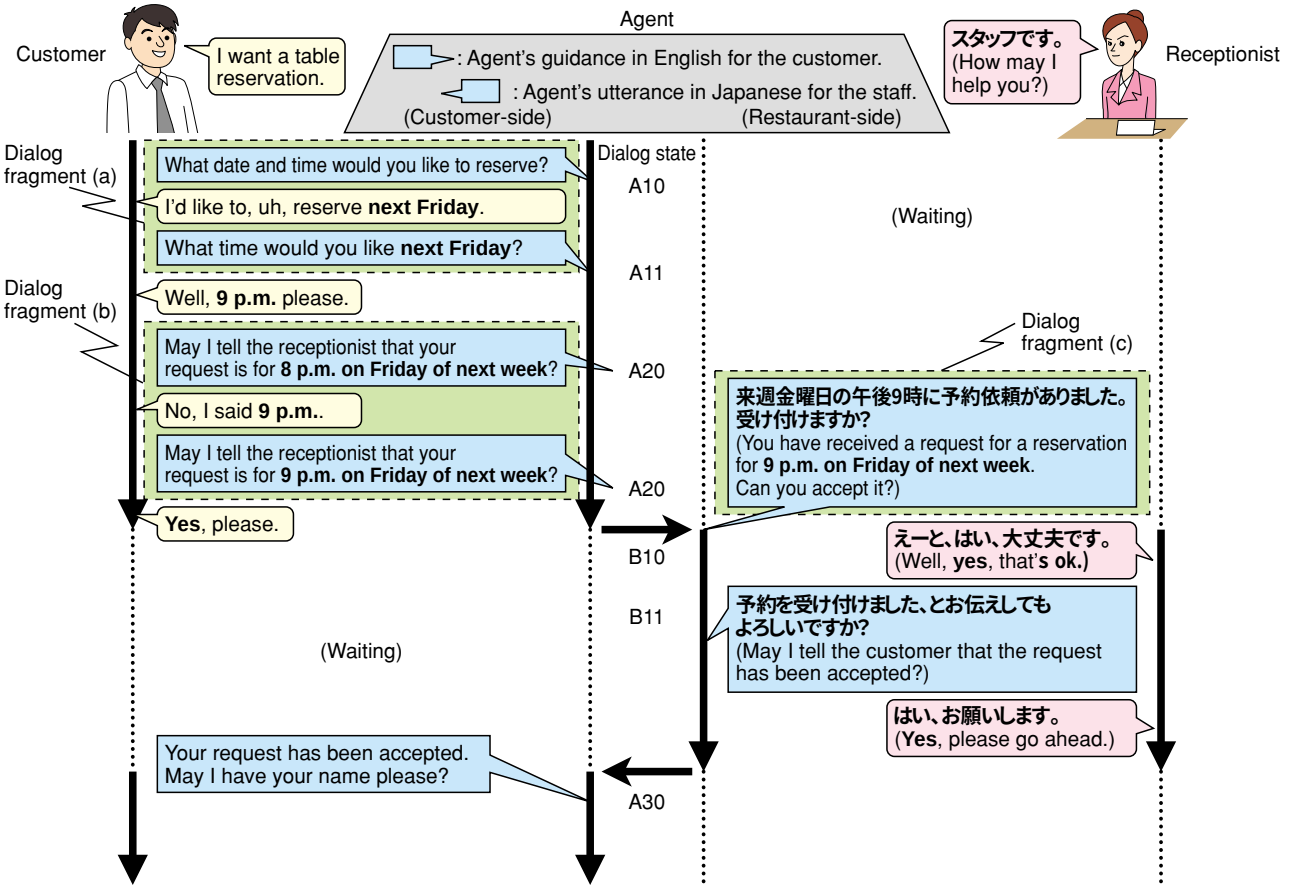

Fig. 2. Example of dialog for making a restaurant reservation.

Table 2. Part of the restaurant-side scenario for making a restaurant reservation.

State	Template sentence of agent guidance	Domain slot	Next state	Submit info.
B10	**[Date]の[Time]に予約依頼がありました。受け付けますか?** (You have received a request for a reservation for **[Time]** on **[Date]**. Can you accept it?)	**[Yes]**	B11	—
		[No]	B12	—
⋮	⋮	⋮	⋮	⋮
B11	**予約を受け付けました、とお伝えしてもよろしいですか?** (May I tell the customer that the request has been accepted?)	**[Yes]**	Submit	A30
		[No]	B20	—
⋮	⋮	⋮	⋮	⋮

ly acquire the date and the time from the customer.

(3) Less machine translation:

It can minimize MT errors. The keywords are acquired from the participant's response to the agent's queries. In the other language, the agent can generate an utterance that conveys the participant's intent by embedding the translated keywords into a sentence template. As shown in dialog fragment (c) in Fig. 2, the translated keywords "*来週金曜日* (Friday of next week)" and "*午後9時* (9 p.m.)" are embedded in a sentence template to convey the customer's request to the Japanese receptionist. This means that the agent does not need to translate the whole utterance, only the keywords. It is easy for MT systems to correctly translate just the keywords with little time, given a known goal.

With these advantages, the agent-mediated architecture can achieve the participants' goal without miscommunication in goal-oriented and cross-lingual communication.

4. Discussion

4.1 Hybrid speech translation

The dialog example in section 3 shows the effectiveness of the interactions between the agent and participants in our architecture. This effectiveness originates in the mediation by an agent that knows the goal of the communication. Traditional speech translation systems convert each utterance and convey the translations without regard for the goal of the communication. The agent-mediated architecture is not only effective for goal-oriented and cross-lingual communication, but it also seems to be effective for communication without an explicitly stated goal, if combined with traditional speech translation systems. For example, we can assume that an agent will monitor each participant's utterance via traditional speech translation systems. When the agent recognizes the goal or sub-goal of the participant, it activates the appropriate dialog scenario and mediates between the participants to achieve that goal.

4.2 Saving labor in describing scenarios

Our architecture requires that an appropriate dialog scenario be provided in advance for each task. This normally entails laborious work. Therefore, we should consider saving labor by preparing dialog scenarios.

One solution is to use authoring tools for a markup language based on XML (extensible markup language) in which dialog scenarios are described in the spoken-dialog component [10]. This tool can reduce the effort required to describe a dialog scenario for the given task. Another solution is to decompose dialog scenarios into sub-scenarios for a small common unit. We plan to make it easy to generate a dialog scenario by combining:

(1) reusable sub-scenarios for a common purpose, such as ones for obtaining the date and name, and
(2) original sub-scenarios specific to a given task.

5. Conclusion

To provide a speech translation system that enables us to accurately achieve the goal of cross-lingual communication with current fallible speech recognition and machine translation technologies, our agent-mediated architecture has two spoken-dialog components coupled via a dialog management component. A spoken-dialog component interacts with a participant in his or her own language to acquire important keywords and then submits them to the other spoken-dialog component until the communication goal is achieved. The dialog management component controls these components and translates the keywords into the other-side's language. Thus, the system with the agent-mediated architecture enables us to accurately achieve the goal of cross-lingual communication.

One of our future challenges will be to achieve real-time cross-lingual communication with as little agent-mediation as possible by integrating this architecture with conventional speech translation systems. Another will be to develop a non-laborious means of speech translation for everyday conversations.

References

[1] R. Gruhn, H. Singer, H. Tsukada, M. Naito, A. Nishio, A. Nakamura, Y. Sagisaka, and S. Nakamura, "Cellular-phone Based Speech-to-Speech Translation System ATR-MATRIX," Proceedings of International Conference on Spoken Language Processing (ICSLP 2000), Vol. IV, pp. 448-451, 2000.

[2] W. Wahlster, "Mobile Speech-to-Speech Translation of Spontaneous Dialogs: Overview of the Final Verbmobil System," Verbmobil: Foundations of Speech-to-Speech Translation, Springer-Verlag; 1st edition, 2000.

[3] K. Yamabana, K. Hanazawa, R. Isotani, S. Osada, A. Okumura, and T. Watanabe, "A Speech Translation System with Mobile Wireless Clients," The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), pp. 133-136, 2003.

[4] B. Zhou, Y. Gao, J. Sorensen, D. Déchelotte, and M. Picheny, "A Hand-held Speech-to-Speech Translation System," Proceedings of 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2003), 2003.

[5] C. Zong, B. Xu, and T. Huang, "Interactive Chinese-to-English Speech Translation Based on Dialogue Management," Proceedings of ACL-02 workshop on Speech-to-Speech Translation: Algorithms and Systems, pp. 61-68, 2002.

[6] E. Sumita, "Corpus-Centered Computation," Proceedings of ACL-02 workshop on Speech-to-Speech Translation: Algorithms and Systems, pp. 1-8, 2002.

[7] H. Blanchon, "A Customizable Interactive Disambiguation Methodology and Two Implementations to Disambiguate French and English Input," Proceedings of International Seminar on Multimodal Interactive Disambiguation (MIDDIM-96), pp. 190-200, 1996.

[8] M. Seligman, "Nine Issues in Speech Translation," Machine Translation, Vol. 15, pp. 149-185, 2000.

[9] A. Waibel, "Interactive Translation of Conversational Speech," IEEE Computer, Vol. 29, No. 7, pp. 41-48, 1996.

[10] J. Hirasawa, T. Amakasu, S. Yamamoto, Y. Yamaguchi, and A. Imamura, "Cyber Attendant System with Spontaneous Speech Interface," NTT Technical Review, Vol. 2, No. 3, pp. 64-69, 2004.

Naoki Asanoma
Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in information science from Waseda University, Tokyo in 1995 and 1997, respectively. In 1997, he joined NTT Communication Science Laboratories. Since then, he has been engaged in R&D of natural language processing. He is a member of the Information Processing Society of Japan (IPSJ) and the Association for Natural Language Processing (NLP).

Setsuo Yamada
Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in electronic engineering from Tokyo Denki University, Tokyo in 1990 and 1992, respectively. In 1992, he joined NTT Network Information Systems Laboratories. Since then, he has been engaged in R&D of natural language processing, especially machine translation. He is a member of IPSJ and NLP.

Osamu Furuse
Senior Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. degree in information engineering, the M.E. degree in information systems, and the Ph.D. degree in information science and electrical engineering from Kyushu University, Fukuoka in 1982, 1984, and 2002, respectively. In 1984, he joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Public Corporation (now NTT). Since then, he has been engaged in R&D of natural language processing. He is a member of IPSJ and NLP.

Masahiro Oku
Senior Research Engineer, Supervisor, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in electronic engineering from Osaka Prefecture University, Sakai, Osaka in 1982 and 1984, respectively, and the Ph.D. degree from Niigata University, Niigata in 2001. In 1984, he joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Public Corporation (now NTT). Since then, he has been engaged in R&D of natural language processing, such as analyzing Japanese idiomatic expressions, detecting and correcting Japanese homophone errors, and creating automated directory assistance systems. He is a member of the Institute of Electronics, Information and Communication Engineers, IPSJ, and NLP.

Akira Kataoka
Engineer, Research and Development Center, NTT West Corporation.

He received the B.E. and M.E. degrees in knowledge-based information engineering from Toyohashi University of Technology, Toyohashi, Aichi in 1998 and 2000, respectively. In 2000, he joined NTT West and moved to NTT Laboratories in August 2000. He was engaged in R&D of machine translation systems until March 2004 when he returned to NTT West. He is a member of NLP.

Yamato Takahashi
Engineer, Portal Business, NTT Resonant Inc.

He received the B.E. and M.E. degrees in information engineering from Niigata University, Niigata in 1992 and 1994, respectively. In 1994, he joined NTT Communication Science Laboratories. Since then, he has been engaged in R&D of machine translation. He is a member of IPSJ and NLP.