3-4 Development of Dialogue Systems "'Kyo-no Hanna' and 'Kyo no Osusume'"

MISU Teruhisa, MIZUKAMI Etsuo, SUGIURA Komei, and IWAHASHI Naoto

We developed dialogue systems of "Kyo-no Hanna" and "Kyono Osusume". Kyo-no Hanna interprets user's spoken queries and provide user appropriate information. Kyo no Osusume can estimates user's potential preferences and then recommends appropriate sightseeing spots that match to the user's preferences. This paper abstracts several technologies used in the systems.

Keywords
Spoken dialog systems, Natural language understanding, Speech recognition, Recommender systems

1 Introduction

Spoken Language Communication Laboratory of Universal Communication Research Institute is aimed at the realization of a society in which system users can use the information system easily, as a natural communication means. We promote the research and development of efficient dialog processing technology, which receives a human-like, natural spoken query, understands/estimates the intent of the query, and then provides appropriate information as a result. We also study a technology which provides information that meets users' queries by estimating users' preferences through interaction with the systems. In order to demonstrate experimentally the results of previous studies and to collect actual data, we released an iPhone application, "AssisTra[*1]" in Jun/2011 (English version[*2] in Mar/2012) which provides tourist with tourism information and a sightseeing spots recommender system, "Kyo no Osusume[*3]" in Oct/2011. This paper outlines the "Spoken dialog processing technology" which is used in these applications. It also outlines the construction method of the recommender system.

2 Kyo-no Hanna

"Kyo-no Hanna", the main feature of AssisTra, is a spoken dialog system which receives a user's natural spoken query and responds to the query with voice and display. Users can check useful information for sightseeing such as sightseeing spots and restaurant information in places, such as Kyoto as shown in Fig. 1 example.

In general, spoken dialog system consists of five modules (element technology). They are automatic speech recognition, spoken language understanding, dialog management, natural language generation, and speech synthesis as shown in Fig. 2.

2.1 Automatic speech recognition, Speech synthesis, and natural language generation

The automatic speech recognition and speech synthesis utilize statistical technique based on Hidden Markov Model. By using a large-scale corpus for learning, they recognize

＊1 http://mastar.jp/assistra/index.html
＊2 http://mastar.jp/kyo-no_hanna/index.html
＊3 http://mastar.jp/kyonoosusume/index.html

natural and continuous spoken sentences and create a synthesized voice close to the human dialog voice. Refer to Subsections **3.2** and **3.3** for details of automatic speech recognition and speech synthesis. Higher speech recognition rate and natural speech synthesis which resembles actual conversation with users is realized by creating a large dialog data of the afore-mentioned for sightseeing in addition to data used for the translation system model learning. Furthermore, we prepared explanatory texts, as texts used in the natural language generation, for sightseeing spots from various viewpoints such as cherry blossoms and autumnal leaf coloration based on the dialog contents of professional guides.

Fig.1 *"Kyo-no Hanna" dialog example*

2.2 Spoken language understanding

Various forms of conversations exist in the human natural dialog system depending on people and on circumstances. For example, when we take a users' intention of trying to find out "ways of accessing sightseeing spots by bus", there are a variety of expressions in the users' dialog such as that shown in Fig. 3. It is not difficult for people to understand the intent of the dialog, but it is important for computers to convert these expressions to an understandable identical symbol (a computer processable language). The spoken language understanding plays this role.

To realize this function, it is important not only to collect users' actual expression, but also to study/develop an accurate spoken language understanding algorithm. We have re-

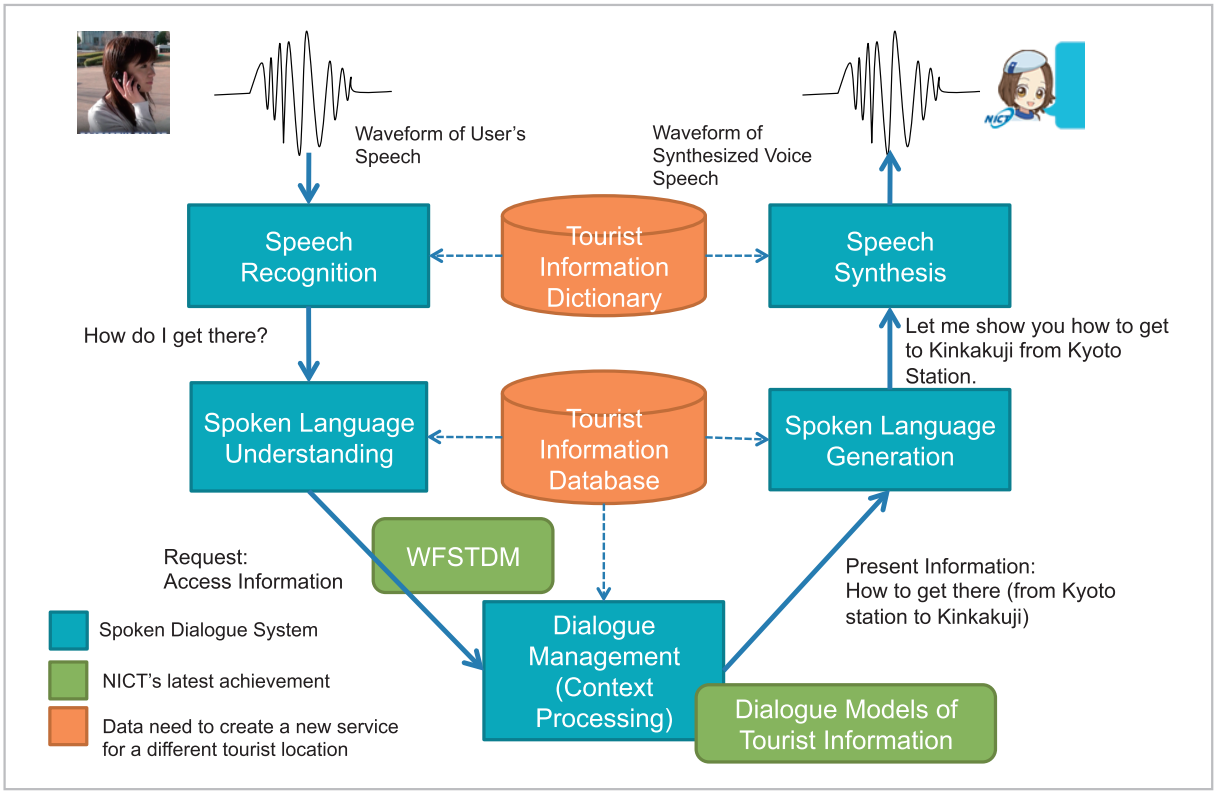

Fig.2 *Spoken dialog system configuration*

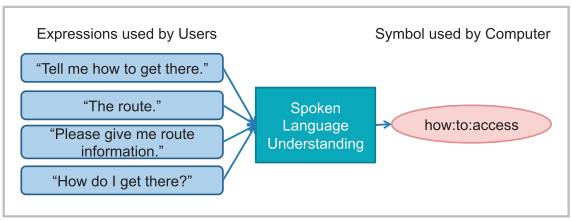

Fig.3 *Examples of spoken language understanding*

corded simulated conversations between professional tour guides and tourists with 150 hours of 300-dialogs to collect actual ways of speaking in conversations [1]. This is a world-class large scale collection of spoken dialog data for a specific situation. Furthermore, we have constructed a prototype of the spoken dialog system, performed a subject experiment, and collected a number of dialog expressions assuming the actual system used. Based on these data, we have realized a fast-and-accurate spoken language understanding by creating a spoken language understanding model with WFST using "WFSTDM: Weighted Finite-State Transducer-based Dialog Manager" [1] which is a spoken language understanding and dialog management framework originally developed by our laboratory.

2.3 Dialog management

Users may expect different system response depending on the circumstances and/or dialog history even when the same dialog is inputted. For example, when the query "access route" is inputted, it is necessary to complement the query with information such as "from where, to where, and by what means of transportation" based on the previous dialog context. It is necessary to decide the response by appropriately compensating requirements which are concealed in the dialog. The dialog processing part plays this role.

Such dialog history processing is severely dependent on usage condition of the dialog system and on the user's purpose of using the system. Therefore, we have created a history processing model for tourist dialog based on the above-mentioned large dialog data that is close to actual users' usage. Then, the dialog history is processed appropriately.

2.4 Summary of Kyo-no Hanna

At present, the application has been published, and the collected log data has being analyzed. However, the system response accuracy is still not enough. We found that variations in human dialog, intention, and ways of speaking is highly varied and complex to be covered by the 150-hour range of learning data. Collection of larger dialog data and improvement of the accuracy of spoken language understanding and dialog history processing is necessary for computers to understand human intention accurately. We will reconstruct each module's model by adding dialog data collected by system operation, and by conducting research into algorithms that understands spoken language flexibly and can manage dialog precisely.

3 Kyo no Osusume

Many tourists collect tourist information orally or from guide books and websites, etc. Based on this information, they decide on places such as Kyoto, Paris or Rome, where they would visit when they go on sightseeing. However, there are too many sightseeing spots in these cities; thus, it takes a lot of time to find out preferred spots. Furthermore, it is difficult to perform a search such as "beautiful garden but not famous spot" using the existing search technology.

Based on these backgrounds, we constructed the sightseeing recommender system, "Kyo no Osusume". Users may get sightseeing spot recommendations easily through touch panel by selecting items such as mood (for healing, for refreshment, etc.), experience, atmosphere, and sightseeing spot character. At the moment, 150 sightseeing spots are currently available.

3.1 Outline of the system

Initial system screen is shown in left picture of Fig. 4. First, users select one category from four of "mood", "experience", "atmo-

sphere", and "spot character". Next, users select evaluation criteria (item) such as shown in the middle picture of Fig. 4. Each sightseeing spot score is calculated based on selected items, and then displayed on the bottom of the screen. When users select a spot, detail information is displayed.

3.2 Extraction of evaluation criteria by Evaluation Grid Method

Two steps were conducted to extract evaluation criteria (item). The first step of the item extraction was conducted by interviewing 24 persons as subjects using the Evaluation Grid Method [1]. Items were classified into four categories, "mood", "experience", "atmosphere", and "spot character" by integrating all the subjects' results and putting them into a common structure using the Evaluation Grid Method. Next, a questionnaire system working on the website (Fig. 5) was constructed to extract items from a number of subjects, and free comment questionnaires were collected from 1,000 subjects. Each item was categorized into the above-mentioned four categories by integrating it based on the synonym. Finally, 137 items such as "world heritage" and "not famous" were acquired.

To utilize the acquired items for recommendations, each item was assumed as an attribute to sightseeing spots, and the attribute score was qualified as a conditional probability. Another questionnaire system working on the website was constructed in the same way as the above-mentioned method, and responds were acquired from 4,000 subjects.

First, we requested subjects to input sightseeing spots "where they have visited and that was satisfactory". Next, we asked them to answer questions with regard to the 137 items by 7 grades. The followings are examples of question items.

1: Very Unlikely, 2: Unlikely, 3: Somewhat Unlikely, 4: Neither Likely Nor Unlikely, 5: Somewhat Likely, 6: Likely, 7: Very Likely

1. Having national treasure level or signature Buddha statue
   ……1 2 3 4 5 6 7
2. Construction and/or interior decoration is rich or characteristic
   ……1 2 3 4 5 6 7
3. Shrines or Temples
   ……1 2 3 4 5 6 7

Each item was binarized, and the conditional probability to the spot was calculated.

Users select items as shown in the middle picture of Fig. 4. Spots scores are calculated by Naive Bayes Method. That is, a product of conditional probability to the selected condition is calculated assuming the value of a certain item does not affect the value of the other items. Preset probability was set as the ratio of users who said "visited and satisfactory". Each sightseeing spot score is calculated and displayed on the bottom of the screen.

4 Summary

Technologies used in "Kyo-no Hanna" and "Kyo no Osusume" released by the Spoken Language Communication Laboratory for smart phone applications was described. The next step of this research and development would be applicable to cases other than sightseeing information domains, such as medical area, educational area with dialog processing and information recommendation technology, and hotel and restaurant search/recommendation.

References

1 K. Ohtake, T. Misu, C Hori, H. Kashioka, and S. Nakamura, "Dialogue acts annotation for NICT Kyoto tour dialogue corpus to construct statistical dialogue systems," In Proc. LREC, 2010.

2 C. Hori, K. Ohtake, T. Misu, H. Kashioka, and S. Nakamura, "Statistical Dialog Management Applied to WFST-based Dialog Systems," In Prof. ICASSP, pp. 4793–4796, 2009.

3 N. Mibayashi, M. Haga, and N. Iwahashi, “Preference Extraction by Grouping EGMfor the Kyoto Tour Guide Dialogue System,” Proc. 4th Spring Annual Meeting of The Japan Society of Kansei Engineering, 14B-01, 2009.

(Accepted June 14, 2012)

***MISU Teruhisa**, Ph.D.*

Researcher, Spoken Language Communication Laboratory, Universal Communication Research Institute

Spoken Language Processing, Spoken Dialogue

***MIZUKAMI Etsuo**, Ph.D.*

Senior Researcher, Spoken Language Communication Laboratory, Universal Communication Research Institute

Evaluation of Dialogue, Communication Science

***SUGIURA Komei**, Ph.D.*

Researcher, Spoken Language Communication Laboratory, Universal Communication Research Institute

Intelligent Robotics, Machine Learning

***IWAHASHI Naoto**, Dr. Eng.*

Senior Researcher, Spoken Language Communication Laboratory, Universal Communication Research Institute

Intelligent Robotics, Multi-Modal Dialog, Human-Robot Interaction