Development of Event-Driven Dialogue System for Social Mobile Robot

Jiang Ridong, Tan Yeow Kee, Li Haizhou, Wong Chern Yuen, Dilip Kumar Limbu
Department of Human Language Technology
Institute for Infocomm Research
1 Fusionopolis Way, #21-01 Connexis, Singapore, 138632
{rjiang, yktan, hli, cywong, dklimbu}@i2r.a-star.edu.sg

Abstract—**This paper is a part of an ongoing project designed to develop a multimodal mobile social robot for office and home environment. An event driven dialogue system architecture is proposed to integrate different components of spoken dialogue system as well as various agents for vision understanding, navigation and radio-frequency identification (RFID) through a number of events and messages. Speech recognition is powered by our in-house developed multilingual, speaker independent phonetic speech recognition engine. A template-based cum rule-based language generation paradigm is proposed to render the interaction so that the dialogue can be evolved based on the context and task domain. Three levels of error recovery strategy are employed to deal with different types of errors. The successful implementation of the proposed system on our experimental social mobile robot is a good showcase toward the integration of spoken language technology with other modalities.**

Keywords-Dialogue management system; event-driven; social mobile robot

I. INTRODUCTION

Spoken language is the medium of communication used first and foremost by humans. With the rapid emergence of various high-tech electronic products and various websites in our daily life, human machine interface becomes extremely important in terms of interaction efficiency and user friendliness. Products equipped with speech interface demonstrate greater competitiveness and attractiveness as speech technology provides intuitive, flexible, natural and inexpensive means of communication with these devices and websites.

In the last decade, speech technology has gained significant advancement. This has resulted in an increasingly widespread use of speech and language technologies in a wide variety of applications, such as voice-operated cell phones, car navigation systems, commercial information retrieval, gaming, education, healthcare, mobile robotics, etc. Some of these systems have successfully introduced into commercial environment [1] [7].

Social mobile robots, as one of the typical and advanced application areas of speech technology, have attracted much more attention in the last few years [2] [6]. This is because we can envision that social robots will affect a broad range of our social activities and impact on a wide range of human lives: bank teller, shop assistants, telephone operators, tour guides, housemaids, elder care nurses, playmates. This in turn brings great challenges to the research on human language technology and multimodal human computer interface.

Spoken dialogue system, as a central interfacing component between human users and the robot, plays important part in the communication. Much effort has been put on the research of spoken dialogue technology in the past thirty years. Different types of dialogue systems have been developed for various application domains. These dialogue systems can be classified into three types according to the methods used to control the flow of the dialogue with the user: Finite State/graph-based; Frame Based/Form based; Plan Based/agent-based. In terms of the dialogue initiative, we can have system directed, user directed and mixed initiative (the initiative is shared) dialogue management systems [1] [3]. These dialogue systems may adopt different language models, possess different kinds of features and apply to different domains. However, there seems less focus on the integration between spoken dialogue system with other modalities and agents. This paper addressed this challenge and proposed an event driven dialogue system for social mobile robot. Besides the description on architecture and key components of spoken dialogue management system, the integration and synchronization between dialogue system and other agents through network message centre and event hub are also discussed

II. SYSTEM CONFIGURATION

As the proposed dialogue system is designed for social mobile robot, besides its interaction capability like other spoken dialogue systems, it is also capable of communicating with operating agents, such as navigation system, vision understanding, RFID, etc. In the following section, we will describe the system architecture and its components of the dialogue system.

A. System Architecture

The dialogue system is composed of speech recognition, dialogue manager, time event generator, language understanding, language generation, speech synthesis, help and error recovery, network message centre and an event hub. The network message centre manages and coordinates the communication and synchronization with other agents - vision understanding, navigation and RFID. The diagram of the system architecture is shown in Fig.1. The event hub

978-0-7695-3571-5/09 $25.00 © 2009 IEEE
DOI 10.1109/GCIS.2009.463

stands in the centre of the whole system to manage different kinds of messages so that various modules and agents are able to cooperate and synchronize seamlessly to perform different functions as a social mobile robot.

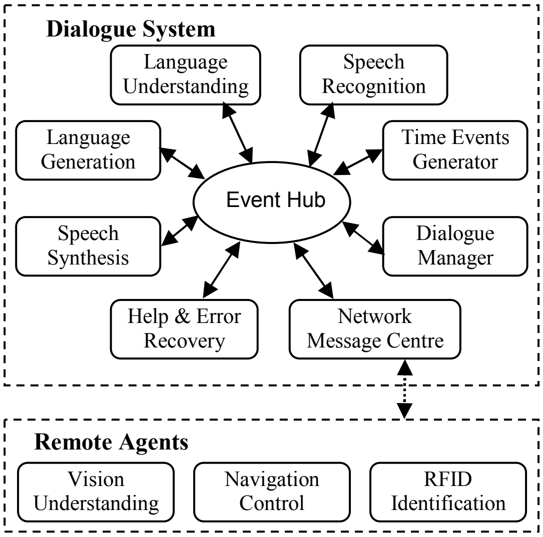

Fig.1. System architecture of event-driven dialogue system

B. Event Hub

Event hub is a relay station of messages. It can do simple message handling and management. Its basic functions include [6]:

- Process received raw messages – The hub receives all messages and interprets the raw messages with pre-defined protocol. The processed messages can be understood by every module and agent.
- Create a new message with necessary information – A new event can be generated and fired by providing task identifier and relevant parameters. Additional information will be automatically added to the event by the hub, such as time stamp, source module of the event, etc.
- Relay messages – After processing the received messages, the hub forwards the messages to relevant components or agents based on the task identifier for necessary actions.

C. Dialogue Manager

Dialogue Manager plays key roles in the dialogue system. It is a module which manages the state of the dialogue, and dialogue strategy. Dialogue manager consists of a script parser and state executor. Different dialogue tasks are specified in text files with a script language. The dialogue is activated and driven by various events which may come from agents through network, or from local modules such as speech recognizer and time event generator. For instance, when a user is detected waving at the robot, the vision understanding agent will send a "user engaged" network message to the event hub. Upon receiving this message, the event hub will relay the message to dialogue manager. Dialogue manager will process the message and trigger an event to language generation and the speech synthesizer to say specific utterance such as "Hi, I can see you".

D. Speech Recognition

Speech recognition is performed based on the ABACUS platform, an in-house developed speaker independent recognition engine [4], a phonetic recognizer for large vocabulary continuous speech recognition. ABACUS is uniquely designed for Asian languages that support recognition of tonal languages, code-switch mixed-lingual and multilingual dialogue implementation. It has been an industry grade spoken dialogue platform deployed in many commercial applications.

ABACUS has two recognition modes: Large Vocabulary Continuous Speech Recognition (LVCSR) and grammar mode. In LVCSR mode, it uses an n-gram statistical language model to support LVCSR. A typical application is speech dictation. In grammar mode, ABACUS supports BNF grammar and VXML grammar definition. It allows the developers to build a dialogue context using simple grammar, such as word list, or complex one as in conversational speech.

The proposed dialogue system was established by grammar-based approach. In different scenarios of dialogue process, different grammars and keywords can be dynamically set and loaded. In this way, the spoken language interaction can be constrained to limited scope and task domain, this helps to improve the speech recognition performance significantly

E. Time Events Generator

Time events generator is a component which generates time events in a pre-defined intervals. As a fully autonomous dialogue system, there are cases that need timers periodically to trigger certain actions. For instance, the network message centre needs to manage the network system to connect with navigation and vision agents. If these connections could not be established when the dialogue system starts, then a timer will be created and activated by the message centre so that dialogue system is able to periodically check the connectivity with these agents. Once these systems are online, the connections can be automatically built up by the timer event. In the dialogue interaction process, some other timers are used to trigger different kinds of events for dialogue state machine, timeout for speech recognition, etc.

F. Language Understanding

Natural language understanding takes the output from speech recognition and analyzes the string of recognized words to derive the meaning embodied in the input speech utterance. Based on the requirements of dialogue system, language understanding can be performed in a way that text string is completely analyzed with computational linguistics syntactically and semantically. If the dialogue system does not require entire sequence of words to be recognized, keyword spotting can be adopted. Keyword spotting techniques deal with recognition of known vocabulary words in unconstrained utterances. It is especially suitable for state-based and system initiative dialogue systems. This method makes the language understanding module more tolerable for incomplete or ill formed speech input, and hence brings robustness to the dialogue system in the sense of input processing.

As a dialogue system for social robot, most of the interaction is command-based and question answering, keyword spotting will be an ideal and efficient approach for language understanding. This paper employs keyword spotting to detect a set of required keywords in the input utterance. The results are then used for the decision of dialogue move.

G. Language Generation

Language generation is constructed by a template-based cum rule-based concept which aims to achieve higher flexibility and adaptability. Template-based language generation approach is widely adopted in many dialogue systems. The templates are designed and selected by the course of interaction. For instance, when the interaction comes to the state of offering a person with his/her favorite drink, a possible template could be:

Hi <Salutation> < LastName> , please help yourself to the <FavoriteDrink> on the tray.

In this template, three fields need to be filled: <Salutation>, <Lastname> and <FavoriteDrink>. The information for these fields is transmitted by RFID agent in the form of message or directly acquired from database according to user's identifier.

The above template-based approach is in the sentence level. It is not enough to achieve more advanced interactions. For the dialogue system for social robot, sophisticated approach is needed to make the communication more natural and vivid. A context-based dialogue evolution can be implemented by rule-based cum template-based approach. Consider the dialogue between robot and human user who possess different interests or characteristics, if the dialogue can be evolved with the knowledge about the user, then the interaction will be more human-like. Following quasi-code shows the rules for an interaction on user's favorites:

```
If #FavoriteSong# is true then
    Converse_Song
If #FavoriteSport# is true then
    Converse_Sport
If #FavoriteMovie# and #FavoriteTV# is true then
    Converse_Movie
Else
    Converse_Generic
EndIf
```

Rules can also be used to generate prompts and help messages. With the combination of rules and templates, the dialogue system is able to render the interaction based on the context and task domain.

H. Speech Synthesis

Speech synthesis produces artificial human speech by providing normal language text. This can be done by concatenating prerecorded pieces of speech or by mathematically modeling the human speech production mechanism. Prerecorded canned speech is suitable for constant messages with far more human-like tone while synthetic speech works well when the output is large amount of texts which are variable and unpredictable. In our system we deployed Microsoft concatenated Text-to-Speech (TTS) Engine through SAPI 5.1 interface. Special SAPI controls can also be inserted along with the input text to change real-time synthesis properties like voice, pitch, word emphasis, speaking rate and volume by synthesis markup which is in standard XML format. In addition, different kinds of events are fired for the purpose of synchronizing to the output speech. The dialogue system is able to produce high quality voice with option on different accents and male or female voice if third part voice packs are installed.

I. Help & Error Recovery

Context-sensitive help and error recovery mechanism is designed to bring more robustness and user friendly features to the interaction. Help information can be provided based on the status of conversation, for instance, the state of the dialogue, the failed times of speech recognition on the same topic. The help information could be standard repeat request "I beg your pardon" and its variation, or further explanation on the current topic. Different rules can be used for help information generation as described in previous section for language generation. This module is independent to other components, it is triggered by help event from different agents.

In the mean time, this module is able to handle certain error situations. As we know, spoke dialogue system may incur different errors which could be caused by speech recognition, language understanding, lack of knowledge on topic and spoken dialogue management itself. In our system, some errors such as speech recognition, network connection with other agents can be detected and handled. There are three levels of recovery strategy when errors occur:

a) To request repeating or provide help information;

b) To advance the dialogue to predefined state based on certain rules and dialogue history;

c) To execute recovery command from control agent.

Strategy a) can be viewed as an effort for error recovery. An example of b) is that when several tries of recognition fail, the dialogue will move to waiting state for conversation:

```
If #NumberOfTry# is 3 then
    WaitToStart
Endif
```

When any errors arise for whatever reason, dialogue system can be recovered by execute command from control agent.

J. Network Message Centre

Network message centre manages network connection with various agents, sends confirmation messages and receives commands such as start conversation and stop conversation, etc. Network message centre is the coordination component between dialogue system and other agents. The typical network message for a service robot is shown in Fig. 2. "WaitToStart" is the state that dialogue stands by and listening to any network messages for conversion. The meaning of network messages is shown as follows:

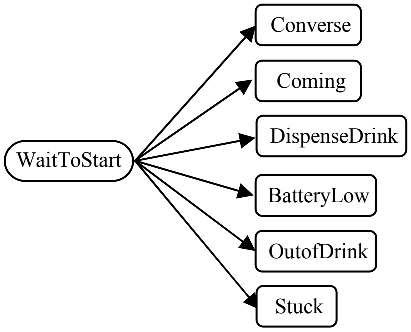

Fig.2. Typical messages for a service robot

a) Converse – message sent by navigation agent for starting conversation on any topic designed for the social robot after robot moved to the user.
b) Coming – when user engaged, the message sent by vision understanding agent and the robot may response by saying "I am coming to you."
c) DispenseDrink – message sent by dialogue manager for drink dispense with possible response such as: "Please help yourself to the drinks on the tray."
d) BatteryLow – message from navigation agent for notifying people that the system battery is low.
e) OutofDrink – message from navigation agent for telling "Please excuse me, my tray is empty, I need to get more drinks."
f) Stuck – message from navigation agent when the robot is blocked by obstacles.

III. Implementation

We have developed an experimental social mobile robot equipped with the proposed event-driven dialogue system and some other technologies such as vision understanding, ultra-wideband localization, radio-frequency identification and navigation control. The robot is able to serve beverage and interact with user on different topics that are specified in a script language. Human user can get engaged by waving his/her hands to the robot, where the input video was captured by stereo camera and sent to the vision understanding agent for processing. Then a message of "Coming" will be sent to the dialogue system to verbally notify the user that he/she has got the robot's attention. The robot starts to approach the user and will stop in front of the user based on the distance estimated by the human-tracking agent. Once it reaches the destination, the speech dialogue will be activated by sending "DispenseDrink" or "Converse" message and the human-robot interaction starts. If the user moves away from the robot, a "DialogStop" message will be sent to the dialogue system by the human-tracking agent. The interaction will stop and the dialogue system will be diverted away to the stand by state.

Experiments were carried out from time to time in research laboratory. Results showed that the proposed event-driven architecture worked effectively in the highly distributed multi-agents environment. Different agents and components were tightly coordinated and synchronized by the event hub and network message centre which successfully make the robot behave and converse more human-like.

IV. Conclusion

This paper presents an event-driven dialogue system for social mobile robot. The proposed system architecture seamlessly integrates different components of spoken dialogue system, as well as various agents for vision understanding, navigation and RFID through a number of events and messages based on certain protocols. The experimental mobile social robot in the home and office environment is a good showcase toward the integration of spoken language technology with other modalities.

The motivation for designing the event driven dialogue system is for social mobile robot. The system architecture contains the necessary components for conventional spoken dialogue system which targets for other task domains. Therefore the dialogue system can be directly applied to similar interaction environment. However, more research and usability study need to be done to improve the system's overall reliability, especially in the real-life scenario. Currently the control strategy of this dialogue system is finite state based, there are inherent advantages and disadvantages for this type of dialogue system. One of the future challenges to us is to enhance the system with more flexible control strategies and incorporate more comprehensive language understanding component to the system.

Acknowledgment

The research described in this paper is the result of collaboration with past and current staff of the Human Language Technology Department and Computer Vision & Image Understanding Department at Institute for Infocomm Research, especially Koh Chin Wei, Ma Bin, Li Liyuan, Hoe Kah Eng, Fachmin Folianto. Their contributions are gratefully acknowledged.

References

[1] MICHAEL F. MCTEAR, "Spoken Dialogue Technology: Enabling the Conversational User Interface", *ACM Computing Surveys,* Vol. 34, No. 1, March 2002, pp. 90–169.

[2] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, "Providing the basis for human-robot interaction: A multi-modal attention system for a mobile robot," *Proceedings International Conference on Multimodal Interfaces.* Vancouver, Canada: ACM, November 2003, pp. 28–35.

[3] D. Bohus et al. 2007. "Olympus: an open-source framework for conversational spoken language interface research". *Proceedings of the Workshop"Bridging the Gap"* at HLT/NAACL 2007.

[4] Haizhou Li, Bin Ma, and Chin-Hui Lee, "A Vector Space Modeling Approach to Spoken Language Identification", *IEEE Transactions on Audio, Speech and Language Processing,* Vol. 15, No. 1, 2007.

[5] Wang, Kuansan. "An Event-Driven Model for Dialogue Systems". ICSLP98 - *Proceedings of the Fifth International*

Conference on Spoken Language Processing, Sydney, Australia, December, pp 393-396.

[6] Terrence Fong, Illah Nourbakhsh, Kerstin Dautenhahn, “A survey of socially interactive robots”. *Robotics and Autonomous Systems* 42 (2003) 143–166.

[7] James, George Ferguson, Nate Blaylock et al, Chester, “Towards a personal medication advisor”, *Journal of Biomedical Informatics* 39 (2006) 500–513.