A Robust Dialogue System with Spontaneous Speech Understanding and Cooperative Response

Toshihiko ITOH, Akihiro DENDA, Satoru KOGURE and Seiichi NAKAGAWA
Department of Information and Computer Sciences
Toyohashi University of Technology
Tenpaku-cho, Toyohashi-shi, Aichi-ken, 441, Japan
E-mail address : {itoh, akihiro, kogure, nakagawa}@slp.tutics.tut.ac.jp

1 Introduction

A spoken dialogue system that can understand spontaneous speech needs to handle extensive range of speech in comparision with the read speech that has been studied so far. The spoken language has looser restriction of the grammar than the written language and has ambiguous phenomena such as interjections, ellipses, inversions, repairs, unknown words and so on. It must be noted that the recognition rate of the speech recognizer is limited by the trade-off between the looseness of linguistic constraints and recognition precision , and that the recognizer may output a sentence as recognition results which human would never say. Therefore, the interpreter that receives recognized sentences must cope not only with spontaneous sentences but also with illegal sentences having recognition errors. Some spoken language systems focus on robust matching to handle ungrammatical utterances and illegal sentences.

The Template Matcher (TM) at the Stanford Research Institute (Jackson *et al.*, 91) instantiates competing templates, each of which seeks to fill its slots with appropriate words and phrases from the utterance. The template with the highest score yields the semantic representation. Carnegie Mellon University's Phoenix (Ward and Young, 93) uses Recursive Transition Network formalism; word patterns correspond to semantic tokens, some of which appear as slots in frame structures. The system fills slots in different frames in parallel, using a form of dynamic programming beam search. The score for frame is the number of input words it accounts for.

Recently many multi-modal systems, which combine speech with touch screen, have been developed. For example, Teil and Bellik developed the tool for drawing coloured geometric objects on a computer display using speech, tactile and a mouse (Teil and Bellik, 91). We also developed a multi-modal dialogue system based on the robust spoken dialogue system.

In Section 2, we present an overview of our spoken dialogue system through multi-modalities. In Section 3, we describe the robust interpreter from errorful speech recognition results and illegal sentences, and in Section 4, we describe the cooperative response generator. In Section 5, we show the results of the evaluation experiments.

2 A Multi-Modal Dialogue System

The domain of our dialogue system is "Mt. Fuji sightseeing guidance (the vocabulary size is 292 words for the recognizer and 948 words for the interpreter, and the test-set word perplexity is 103)", The dialogue system is composed of 4 parts: Input by speech recognizer and touch screen, graphical user interface, interpreter, and response generator. The latter two parts are described in Sections 3 and 4.

2.1 Spontaneous Speech Recognizer

The speech recognizer uses a frame-synchronous one pass Viterbi algorithm and Earley like parser for context-free grammar, while using HMMs as syllable units. Input speech is analyzed by the following conditions :

Sampling frequency :	12kHz
Hamming window size :	21.33ms (256 samples)
Frame period :	8ms
LPC analysis :	14th order
Feature parameter :	10 LPC Mel-cepstram coefficients and regression coefficients (ΔCEP)

The acoustic models consist of 113 syllable based HMMs, which have 5 states, 4 Gaussian densities and 4 discrete duration distributions. The speaker-independent HMMs were adapted to the test speaker using 20 utterances for the adaptation. The grammar used in our speech recognizer is represented by a context-free grammar which describes the syntactic and semantic information.

Our recognizer integrates the acoustic process with linguistic process directly without the phrase or word lattice. We could say that this architecture is better for not only cooperatively read speech but spontaneous speech rather than hierarchical architectures interleaved with phrase lattice (Kai and Nakagawa, 95). Furthermore, the recognizer processes interjections and restarts based on an unknown word processing technique. The unknown word processing part uses HMM's likelihood scores for arbitrary syllable sequences.

A context free grammar is made to be able to accept sentences with omitted post-positions and inversion of word in order to recognize spontaneous speech. We assume that the interjections and

restarts occur at the phrase boundaries. Thus, our speech recognizer for read speech was improved to deal with spontaneous speech.

2.2 Touch screen (pointing device)

The touch panel used here is an electrostatic type produced by Nissya International System Inc. and the resolution is 1024 × 1024 points. This panel is attached on the 21 inch display of SPARC-10, which has coordinate axes of 1152 × 900 and a transmission speed of 180 points/sec.

The input by touch screen is used to designate the location of map around Mt.Fuji (which is a main location related to our task) on the display or to select the desired item from the menu which consists of the set of items responded by a speech synthesizer. The response through the speech synthesizer is convenient, however, user cannot memorize the content when the content includes many items. Therefore, we use the display output (map and menu) as well as speech synthesis for the response. User can only use the positioning/selecting input and speech input at the same time. For example, user can utter "Is here ... ?" while positioning the location or menu. In this case, system regard the demonstartive "here" as a keyword that user has positioned/selected.

2.3 Graphical User Interface

On man-machine communication, user wants to know his or machine situation what information he gets from the dialogue or how machine interprets/understands his utterances, as well as the speech recognition result. Therefore our system displays the history of dialogue. This function helps to eliminate user uneasiness. Figure 1 illustrates an example of map, menu and history. A multi-modal response algorithm is very simple, because the system is sure to respond to user through speech synthesizer and if the system is possible to respond through graphical information, the system does use these.

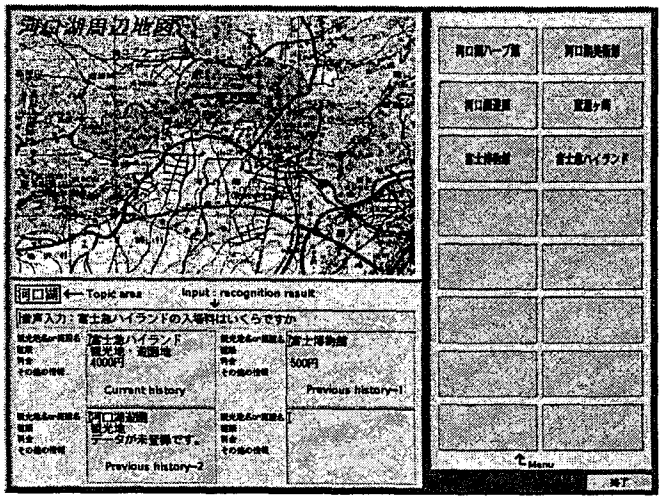

Figure 1: An Example of Map, Menu, and History for an Input Utterance
(Input : How much is the entrance fee for Fujikyu-Highland ?)

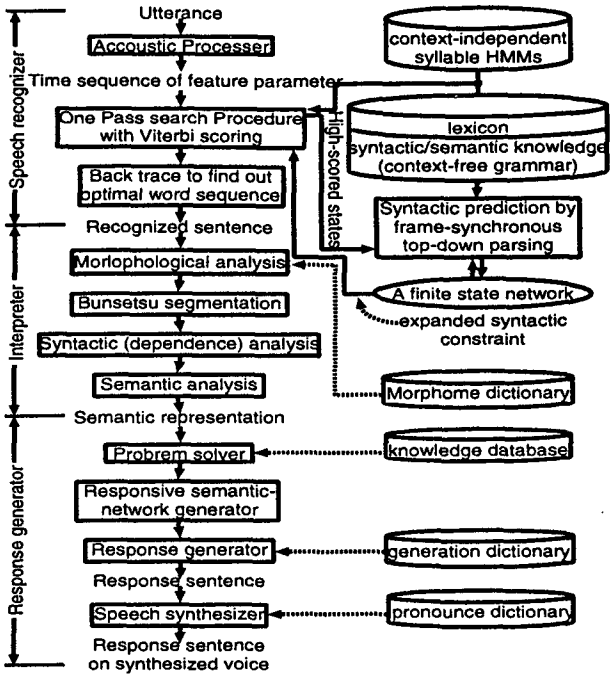

Figure 2: Spoken Dialogue System

3 The Interpreter

3.1 Processing illegal Utterances

Whole process is carried out as below:

1. The steps in the following process are carried out one by one. When one of the steps succeeds, go to process 2. If all of the processes fail, go to process 4.
   (a) syntax and semantics analysis for legal sentence without omission of post-positions and inversion of word order.
   (b) syntax and semantics analysis for sentence including omission of post-positions.
   (c) syntax and semantics analysis for sentence including omission of post-positions and inversion of word order.
   (d) syntax and semantics analysis for sentence including invalid (misrecognized) post-positions and inversion of word order.
2. Fundamental contextual processing is performed.
   (a) Replace demonstrative word with adequate words registered for a demonstrative word database
   (b) Unify different semantic networks using default knowledges, which are considered to be semantically equivalent to each other (processing for semantically omissions).
3. Semantic representation of the sentence is checked using contextual knowledge (we call it filtering hereafter).
   (a) correct case: Output the semantic representation of the analysis result (end of analysis).

(b) incorrect case: If there are some heuristics for correcting, apply them to the semantic representation. The corrected semantic representation is the result of analysis (end of analysis). If there aren't any applicable heuristics, go to process 4.

4. Keyword analysis (later mentioned) is performed by using a partial result of the analysis.

First, the interpreter assumes that there are no omissions and inversions in the sentence(1-a). Second, when the analysis fails, the interpreter uses the heuristics which enable to recover about 90% of inversions and post-position omissions(Yamamoto *et al.*, 92)(1-b,c). Furthermore, when the interpreter fails the analysis using the heuristics, it assumes that the post-position is wrong. Post-positions assumed to be wrong are ignored and the correct post-position is guessed using above heuristics(1-d). The interpreter gives the priority to the interpretation where the number of post-position assumed to be wrong is a few as possible.

Human agents can recover illegal sentences by using general syntactical knowledge and/or contextual knowledge. To do this process by computer, we realized a filtering process(3-b). Contextually disallowable semantic representations are registered as filters. This process has 2 functions. One is to block semantic networks including the same as the registered networks for wrong patterns. The other is to modify networks so that they can be accepted as semantically correct. If the input pattern matches with one of the registered patterns, its semantic representation is rejected, and the correction procedure is applied if possible. The patterns are specified as semantic representations including variables, and the matching algorithm works a unification-like.

When no network is generated at this stage, the interpreter checks the sentence using keyword based method(4). The interpreter has several dozens of template networks which have semantic conditions on some nodes. If one of them is satisfied by some words in the sentence, it is accepted as the corresponding semantic network.

4 The Cooperative Response Generator

Dialogue system through natural language must be designed so that it can cooperatively response to users. For example, if a user's query doesn't have enough conditions/information to answer the question by sysytem, or if there is much retrieved information from the knowledge database for user's question, the dialogue manager queries the user to get necessary conditions or to select the candidate, respectively. Further, if the system can't retrieve any information related to the user's question, the generator proposes an alternative plan. Based on these considerations, we developed a cooperative response generator in the dialogue system.

The response generator is composed of dialogue manager, intention(focus) analyzer, problem solver, knowledge databases, and response sentence generator as shown in Figure 2 (lower part).

Firstly, the dialogue manager receives a semantic representation (that is,semantic network) through the semantic interpreter for the user's utterance. The dialogue manager is a component which carries out some operations such as dialogue management, control of contextual information and query to users. Secondly, to get intention for managing dialogues, the dialogue manager passes semantic network to intention(al) analyzer which extracts a dialogue intention and conditions/information of a user's query.

Then, the dialogue manager decides a flow of dialogue using the intention that is sent back from the intention analyzer and acquires available information from dialogue history as contextual information. Thirdly, the dialogue manager passes a semantic network and contextual information to problem solver to retrieve any information from the knowledge database. Further, if the problem solver can't retrieve any information related to the user's question, the problem solver proposes an alternative plan (information) by changing a part of conditions of usr's query and send it back to dialgoue manager.

Then the dialogue manager counts a number of retrieved information. If there is much retrieved information from the knowledge database for user's question, the dialogue manager queries further conditions to the user to select the information. If the number of these is adequate, the dialogue manager gives a semantic network and retrieved information to the response sentence generator.

Finally, the response sentence generator decides a response form from the received inputs and then forms response sentence networks according to this form. After this process was finished, the response sentence generator converts these networks into response sentences.

5 Evaluation Experiment

5.1 Overview

In order to evaluate our dialogue system with the multi-modal interfaces, we investigated its performance through the evaluation experiments, paying attention to "usefulness of our system".

We gave a task of making some plans of Mt.Fuji sightseeing to 10 users[A ... J] (6 users where evaluation of language processing part) who did not know about this system[novises] in advance. The number of items that user should fill in using our system in this experiment is eight: "Where to go" and "What to do" in first day and second day, and "Where to stay", "Kind of accommodation", "Accommodation name", and "Accommodation fee" in first night. We explained this dialogue system to them and asked them to speak to the system freely and spontaneously.

And we gave three dialogue modes to every subjects, as shown in below :

mode-A Using only speech input and output (our conventional system)

mode-B Using speech input and multi-modal output (graphical output on display and speech output)

mode-C Using multi-modal input and output (input : speech and using touch screen, output : speech and graphic on display)

Users used three systems on-line mode at the computer room.

In this experiment, the performances (recognition / comprehension rate, dialogue time, number of utterances) of three systems were not seen explicit differences, because the system is imperfect.

5.2 Evaluation of the language processing part through the experimental result

Table 1 shows the performance of our system through experiments using mode-A system, which investigated the performance of the language processing parts.

The column of "Speech input" is the result that experiments was done in practice. And the column of "Text input" is the perforamnce of our system, when system inputted a transcription of user's utterances that the recognition rate of the speech recognizer was assumed as 100%. "Semicorrect Recog" means the recognition rate that permitted some recognition errors of particles. "Data presentation" is the rate that the system offered the valuable information to user. "System query" is the rate that the system queried the user to get necessary conditions and to select the information. "Alternative plan" is the rate that the system proposed the alternative plan. "Correct response" is the sum of "Data presentation", "System query", "Alternative plan" and rate that the intepreter was unsuccessful in generating a semantic network. "Retrieval failure" is the rate that the system could not offer the valuable information to user although the interpreter has been successful in generating a semantic network.

The number of total utterances was 101. 81 out of 101 were acceptable by the grammar of the recognizer. 12 unacceptable out of 20 utterances were caused by unknown words, so we considered that it was very important to solve the unknown word problem. And, 8 out of 20 were not acceptable by the grammar. The recognition rate of the speech recognizer on the spontaneous speech was 20.8%. In the speech input, the system unterstood about 55% of the all utterances and offered the available information to user about 55% (42.6%+9.0%+3.0%). And in the text input, these rates were 90% and 80%, respectively. These rates show that the language processing part worked well.

Table 1: Evaluation results

Evaluation	Speech input sentence(%)	Text input sentences(%)
Subjects(users)	6 users	
Utterances	101(100%)	
Correct recognition	21(20.8%)	—
Semicorrect Recog	56(55.4%)	—
Interpretation	56(55.4%)	90(89.1%)
Correct response	81(80.2%)	87(86.1%)
Data presentation	43(42.6%)	64(63.4%)
System query	9(9.0%)	12(12.0%)
Alternative plan	3(3.0%)	5(5.0%)
Retrieval failure	4(4.0%)	9(8.9%)

6 Conclusion

We developed the robust interpreter that can accept not only spontaneous speech but also misrecognized sentences. The interpreter was implemented to our dialogue system for spontaneous speech which worked in the task domain of "Mt.Fuji sightseeing guidance". Further more, based on that dialog system through natural language must be designed so that it can cooperatively response to users, we devloped a cooperative response generator in the dialogue system. This dialogue system was integrated with a touch screen input method. Experiments showed that our interpretation mechanism is suitable for understanding the recognition result of spontaneous speech. And we found that the multimodal interface with spontaneous speech and touch screen was user-friendly.

References

E.Jackson, J.Bear, R.Moore, and A.Podlozny: "A template matcher for robust NL interpretation" Proc. Speech and Natural Language Workshop, Morgan Kaufmann Inc., pp.190-194, Feb.19-22 (1991).

W.Ward and S.Young: "Flexible use of semantic constraints in speech recognition," Proc. Int. Conf. Acoustics, Speech & Signal Process, vol. II, pp.49-50, Minneapolis (1993).

D.Teil and Y.Bellik : Multimodal dialogue interface on a workstation, Venaco Workshop and ETRW on "The structure of multimodal dialogue", Maratea (1991).

A.Kai and S.Nakagawa : "Investigation on unknown word processing and strategies for spontaneous speech understanding", Proceedings of EUROSPEECH 95, pp.2095-2098 (1995).

M.Yamamoto, S.Kobayashi, S.Nakagawa: "An analysis and parsing method of the omission of post-position and inversion on japanese spoken sentence in dialog", Transactions of Information Processing Society of Japan Vol.33, No.11, pp.1322-1330(1992), in Japanese.

T.Itoh, M.Hidano, M.Yamamoto and S.Nakagawa : "Spontaneous Speech Understanding for a Robust Dialogue System", Proceeding of Natural Language Processing Pacific Rim Symposium '95, Volume 2, pp.538-543 (1995).

A.Denda, T.Itoh, and S.Nakagawa : "A Robust Dialogue System with Spontaneous Speech and Touch Screen", Proceeding of the First International Conference on Multumodal Interface '96, pp.144-151 (1996).