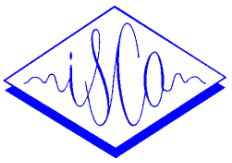

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A New Dialogue Control Method Based on Human Listening Process to Construct an Interface for Ascertaining a User's Inputs

Masanobu Higashida
ATR International
2-2-2, Hikaridai, Seikacho, Sorakugun,
Kyoto-fu, Japan
+81 774 95 1170
higasida@ctr.atr.co.jp

Kumiko Ohmori
NTT Information Sharing Platform Labs
1-1, Hikari-no-oka, Yokosuka-Shi
Kanagawa-ken, Japan
+81 468 593745
kumiko@isl.ntt.co.jp

ABSTRACT

This paper describes a new dialogue control method that utilizes new recognition processes called "Presupposition-type Recognition" and "Pretense-type Recognition" that we propose based on human dialogue analysis. This method provides users with stress-free voice input through real-time responses, comprising a naturally controlled dialogue to obtain information in order to winnow candidates comprehensively.

1. INTRODUCTION

Recent advances in speech recognition technology have accelerated system development using voice input as a man-machine interface, such as GALAXY developed by Goodien (1994) and the PEGASUS method established by Zue (1994).

In addition to user benefits such as ease of use and a method that does not require any practice, using a verbal interface is expected to bring a plethora of advantages to service providers, such as reducing man-hours and the ability to provide around-the-clock services. To apply voice technology to the front-end of various services, the main premise of this technology is to accurately ascertain the user's intent.

However, in practical situations, various limitations hinder the construction of commercial systems that deal with speech recognition technology. For example, systems that handle a huge number of target words mostly suffer from such problems as serious performance deterioration or decreased recognition accuracy. Due to these, the voice input interface is criticized as being "circumlocutory", "having frequent misinterpretations", "having an excessively long processing time", etc., which results in a loss of user satisfaction as described in Sagayama (1994). Therefore, it is only used in a limited number of fields as stated in Zue (1997).

This paper proposes a new dialogue control method that ascertains the user's inputs while taking the above difficulties into consideration. The experimental results confirm that the method is highly efficient in ascertaining the user's intent while not making the user feel stressed, and show that the method can adequately handle speech recognition errors in the dialogue discourse.

2. EXISTING USER-INTENT CLARIFICATION METHOD

2.1. Speech Recognition Technology

When constructing a system that uses existing speech recognition technologies, we must keep in mind two points.

- The number of words that can be recognized within the time a natural dialogue can be held is limited.
- Recognition accuracy is very dependent on the usage environment and the conditions, and we cannot always obtain the best processing performance and functions from the recognition engine.

For these reasons, even if only the keywords are considered (nouns, verbs, etc.), the number of target words for recognition must be limited to maintain a degree of naturalness in a dialogue and to guarantee the recognition rate to a certain extent. Therefore, in a practical service where a large vocabulary must be used as the recognition target, the following procedures were adopted in order to obtain the target words when using a voice interface.

2.2. Existing Dialogue Control Method

The existing interface in many cases uses a dialogue control system that is favorable to the system, as described below.

1. By hierarchically classifying the attributes necessary to ascertain the user's intent, a large vocabulary is classified such that the number of words in the nodes of each level becomes small. The user inputs are ascertained by iteratively confirming words in sequential order from the top level down and winnowing the lower levels.
2. Recognition results are presented to the user based on the score and a confirmation cycle is performed repeatedly until the correct answer is confirmed. By definitely ascertaining

each attribute value, the number of target words in the next lower level is reduced.

For these reasons, the existing interface has several problems:

1. To identify the user's inputs, the confirmation cycle must be repeated at least as many times as the number of levels of the hierarchy.
2. The process does not advance until the correct word that was spoken by a user is ascertained.

These are the reasons why users experience irritations such as redundant dialogues.

3. HUMAN STRATEGIES TO CONTROL DIALOGUES

3.1. Hypothesis on Human Recognition Process

Even in dialogues between people, depending on the conditions surrounding a conversation, it may be hard to clearly understand what was said or easy to mistake what was said for something very similar. When people are faced with this type of problem, people seem to pursue a related dialogue to obtain supporting information to clarify the information that was exchanged.

By analyzing dozens of recorded dialogues between operators and users in a telephone service such as directory assistance or customer service, we formed a hypothesis to explain operators' behaviors in the process of ascertaining users' inputs. This hypothesis consists of three dialogue control strategies.

Hypothesis: Human operators have two kinds of listening strategies that are constructed in a recursive manner in a practical operation.

(**Strategy 1)** In a conversation, the operator usually considers a small number of words when a word first spoken by a user is recognized, even though a large number of words should be equally considered as possible candidates. These words are selected and collected in the operator's memory by preference, based on the bias of the spoken words or on their relation to daily life.

(Strategy 2) When Strategy 1 fails for some reason, the operator tries to obtain additional information to identify the user's inputs, not letting the user know that the operator could not fully understand what was said.

(Strategy 3) Even in the process of interaction in Strategy 2, Strategy 1 is adopted if the number of target words is large. When this strategy fails again, Strategy 2 is recursively called upon until enough information is obtained to ascertain the user's inputs.

3.2. Evidence from Experiments

To establish this hypothesis, a series of questions were provided to operators through interviews. This study showed interesting results to support the hypothesis described above and to help make a computational model to simulate human dialogues.

(1) Words prepared for a dialogue:

A list containing all Japanese address segments (4,100) at the second level next to prefectures was presented to check the operators' knowledge level. The average number of segments that operators know well is about 300 and the rest of the segments were unfamiliar to them. Regarding family names, only 900 out of 160,000 diversities in pronunciation are well known to operators. These show that even operators consider only a small portion of the total target words. These observations seem to support the first strategy.

(2) Status after hearing a word spoken by the user:

Operators' responses can be categorized into four classes.

(a) The spoken word is clearly heard and identified without ambiguity.

(b) Though the spoken word is recognized, multiple candidates still remain.

(c) Though the spoken word is recognized as a series of syllables, it is not a familiar one.

(d) The spoken word fails to be recognized for some reason, such as the speaker's voice conditions, noises in the speaker's environment or carelessness of the listener.

These results can be used in the modeling of results in a recognition engine.

(3) Policies when operators feel uncertainty or failure in listening:

To cope with the problems in the cases of (b), (c) and (d), operators showed the following policies.

(b1) Obtain additional information to dissolve the ambiguities, by a Yes/No question or a question to solicit the user to input other information.

(c1) Ask another question to the user to winnow the retrieval space in the operator's mind.

(d1) Ask the user to input the same word again.

We believe that the hypothesis concerning dialogues controlled by human operators described in the head of this section can be clarified by the evidence observed from questionnaires and interviews with operators.

Interestingly, the explanation in (1) reminds us of the requirement for recognition engines now available to have limitations in the number of recognition target words if good precision with good performance is requested in practical systems. And moreover, the

conditions listed in (2) connote a mapping possibility to the recognition results in recognizers.

4. PROPOSED DIALOGUE CONTROL

4.1. Computational dialogue model using a recognition engine

Based on the analysis in Section 3, we tried to make a computational dialogue control model that efficiently manages interactions with users using voice recognition technologies. The three strategies described in section 3 were named after their characteristics so that we can easily understand the concepts from the name.

Strategy 1: Presupposition-type Recognition (RSR)

Strategy 2: Pretense-type Recognition (PTR)

Strategy 3: Recursive Dialogue Control (RDC)

The structure of these strategies is illustrated in Fig. 1.

Recognition processes based on PSR and PTR strategies are connected in a cascade manner to implement the RDC strategy.

A PSR will be initiated at the beginning of the dialogue by selecting an appropriate attribute. If PSR is successful in identifying a word spoken by the user, or if the target word set that should be handled in the system is small enough to be processed in a real-time interval, PTR will not be initiated. If PSR fails for some reason, PTR is complementarily initiated to support the clarification process or to back up the recognition for those words not included in the PSR target words.

The attribute introduced in the initial stage of the dialogue is normally selected by the statistical analysis of the user's behaviors or the operator's intentions to guide the procedure.

4.2. Modeling of the PSR Process

The PSR process is characterized by using a recognition engine with a limited set of target words. The important parts of this model are the selection of the words included in the target word set and the categorization of the recognition results.

The limited target words are preferentially selected based on pre-estimated attribute values using the bias of the user's spoken frequency or a statistical list of attribute values in order of frequency. The words are limited to a number that can be

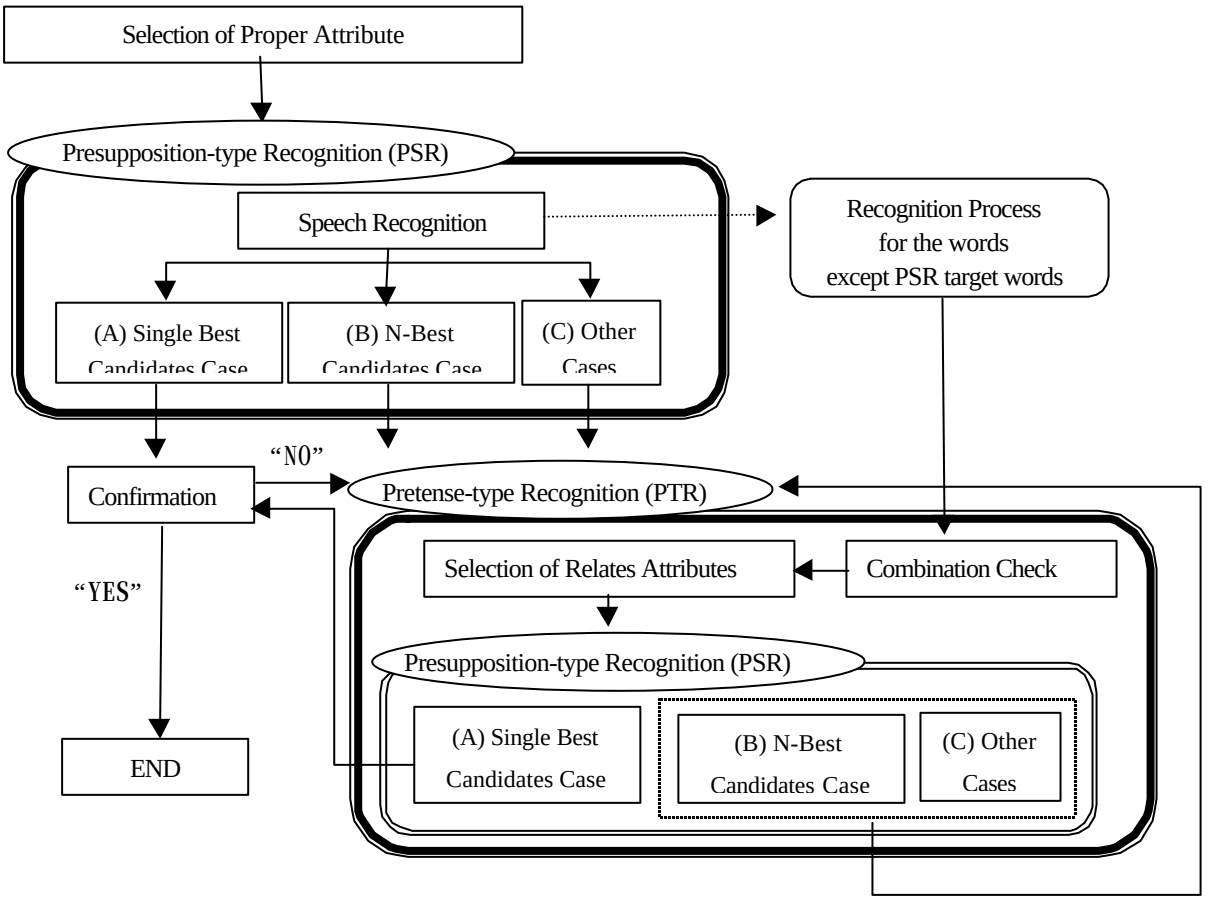

Fig1: Proposed dialogue process model

processed within a specified time to maintain the naturalness of the dialogue. This time is defined depending on the recognizer performance.

Recognition results are generally provided with numerical scores. Results will be categorized into three types by comparing the interview result (2) in section 3 and the recognition likelihood of the candidates in the top levels.

Type A - Single Best Candidate Case: Item (a) in result (2) corresponds to this case. A top candidate with a high score and with no close runner-ups exists in the list of candidates. Because the recognizer provides confidence to identify the result, only a confirmation process will be needed in this case.

Type B - N-best Candidates Case: Item (b) in result (2) can be considered to fall into this type. Top N candidates form an unrivaled group. The correct answer is most likely included in this group. The number is most often set to three or less to facilitate dialogue control in a practical case.

Type C - Other Case: In this case, no special suggestion from the recognizer is provided. The recognizer only gives a list of candidates having continuous scores. Items (c) and (d) seem to fall into this type.

When the PSR process gives Type B or Type C results, additional interactions with the user will become necessary to identify the result.

To implement this process, a real-time response can be realized for a user that inputs a word among the PSR target words, even if the system must handle a huge number of words. Because a spoken word is mostly recognized as one of the few top candidates after a quick recognition process, this process can cover most of the users by providing them with a tuned-up and high performance recognizer.

4.3. Modeling of the PTR Process

The PTR process is initiated only when the first PSR process fails to identify a spoken word. In PTR, another PSR process is initiated. If the target word set is small enough to be processed in real-time, the recognition process will be terminated at this level. When the target words are still large in number, a third PSR process in the second PTR process might be initiated.

The purpose of PTR is to obtain additional information from the user in three ways:

(T1) to ask the user to repeat the word because the recognizer seems to have low scores for all candidates,

(T2) to ask the user to answer Yes/No questions to resolve ambiguities in recognized candidates,

(T3) to ask the user to input another value for the presented attribute selected by the system. The information obtained from the last one is used in two ways; to winnow the target words to a smaller set if an attribute value at a higher level is obtained, and to select the correct answer by testing mutual connectivity among the two kinds of attribute value candidates provided by the recognizer.

5. EXPECTED EFFECTIVENESS FROM THE PROPOSED METHOD

5.1. Comparison with the Existing Method

Here, a belief comparison of a human dialogue is made with one conducted by a computer employing voice recognition based on the decrease of information entropy obtained from the user in the discourse. As an example, consider ascertaining the large sections of municipalities in Japan, such as "Chiyoda-ku", "Yokohama-shi", or "Soraku-gun", from the address segments which can be hierarchically organized. The 180,000 address segments are classified one by one by prefecture (47 words) and municipality (4,100 words) in a hierarchical manner. The existing voice input system as in Chiristopher (1993), which decreases the information entropy, is shown in Fig. 2(i) based on roughly estimated values.

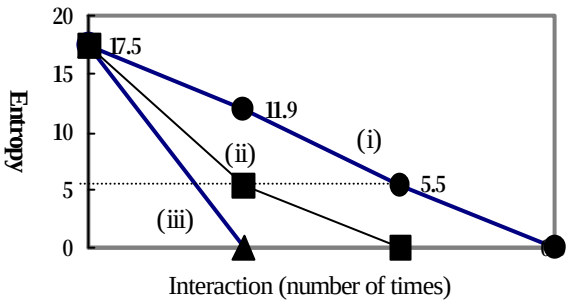

Fig.2: Decrease process of information entropy

At the beginning of a dialogue, the information entropy is 17.5 bits. In the existing method, prefectures are defined as having an information entropy of 5.5 bits, and the municipalities are defined as having an information entropy of 11.9 bits. These decrease stepwise until the information is ascertained. The existing method is slow and is considered to be a source of user dissatisfaction. If we could trace the human dialogue, the new information-entropy decreasing process shown in Figs. 2(ii) and 2(iii) is implemented.

Namely, by enabling input from the large sections of the 4,100 municipalities or even from the lower level of the 180,000 address segments, we aim to achieve comparably faster and more accurate reduction in information entropy.

6 CONCLUDING REMARKS

In this paper we proposed a new dialogue control method based on an analysis of the listening process of human operators. Development of three prototype systems is now underway, targeting the ascertainment of addresses, personal names and ticket reservations.

The time spent for a dialogue and number of interactions required to complete a task will be measured after their development. Questionnaires to the subjects will also be done to study the users' customer satisfaction (CS) level. Comparison with the existing method is expected to show that the proposed method improves performance and efficiency in dialogues to some extent.

7. ACKNOWLEDGMENTS

The authors thank Mr. Takaaki Matsumoto, project manager, and Dr. Jun Sekine, group leader, for the opportunity to conduct this research. The authors also thank Dr. Koji Dosaka and Ms. Yukiko Nakano for their helpful advice.

8. REFERENCES

1. Chiristopher S. (1993) "Voice Communication With Computers," Van Nostrand Reinhold, A Division of Wadsworth, Inc.
2. Goodien D., et al. (1994) "GALAXY: A human-language interface to on-line travel information," *Proc. ISCLP '94*, pp. 707-710.
3. Sagayama, S. (1994) "Why Is Speech Recognition Not Widely Used? How Can It Be Used More Often? ," *IPSJ SIG Note, 94-SLP-1*, pp. 23-30.
4. Zue V., et al. (1994) "PEGASUS: A spoken dialogue interface for on-line air travel planning," *Speech Commun. 15*, pp. 331-340.
5. Zue, V. (1997) "Conversational interfaces: advances and challenges," *Proc. Eurospeech*, KN-9-KN-18.