Efficient Spoken Dialogue Control Depending on the Speech Recognition Rate and System's Database

Kohji Dohsaka, Norihito Yasuda, Kiyoaki Aikawa

NTT Communication Science Laboratories, NTT Corporation, Japan
{dohsaka, yasuda}@atom.brl.ntt.co.jp, aik@idea.brl.ntt.co.jp

Abstract

We present dialogue control methods (the dual-cost method and the trial dual-cost method) that enable a spoken dialogue system to convey information to the user in as short a dialogue as possible depending on the speech recognition rate and the content of its database. Both methods control a dialogue so as to minimize the sum of two costs: the confirmation cost (C-cost) and the information transfer cost (I-cost). The C-cost is the length of a subdialogue for confirming a user query, and the I-cost is the length of a system response generated after the confirmations. The dual-cost method can avoid the unnecessary confirmations that are inevitable in conventional methods. The trial dual-cost method is an improved version of the dual-cost method. Whereas the dual-cost method has the limitation that it generates a system response based on only the content of a query that the user has acknowledged in the confirmation subdialogue, the trial dual-cost method does not. Dialogue experiments prove that the trial dual-cost method outperforms the dual-cost method and that both methods outperform conventional ones.

1. Introduction

Spoken dialogue systems perform tasks like information retrieval and making reservations through speech communication with their human users. Consider a system with a database that can handle several types of user query. Due to speech recognition errors, a system carries out a confirmation subdialogue whose purpose is to determine the content of a query. In this subdialogue, the system confirms whether the content of the currently recognized query is correct or not. If correct, the user makes an acknowledgement like "Yes". Otherwise, the user corrects the system's misunderstanding. Although confirmations are helpful in avoiding misunderstandings, continual confirmations by the system interfere with the smooth flow of the dialogue. Therefore, it is desirable that the system should avoid unnecessary confirmations.

Typical unnecessary confirmations occur when a system makes confirmations irrespective of the content of its database [2]. For example, consider a system for weather information retrieval. Suppose that the system recognizes that the user has made a query as to whether a heavy rain warning has been issued for Tokyo and that no warning has been issued anywhere is stored in the system's database. In this situation, the system does not have to confirm the place; it only has to confirm the type of information since it can respond that there are no warnings issued anywhere. The confirmation of the place is unnecessary since its omission does not severely affect the length of the system's response. The place confirmation makes the whole dialogue long and there is a risk of misrecognition during that confirmation. In order to avoid such unnecessary confirmations, the dialogue control method must try to reduce the total length of a dialogue including a system response according to what is in the system's database.

Previous work on dialogue control [1, 4, 5, 6] have focused on reducing the length of a confirmation subdialogue and does not pay much attention to the effect of the database content on the total dialogue length. Therefore, if those methods are directly applied, it is difficult to avoid unnecessary confirmations. We have proposed a dialogue control method that works on the basis of the content of the system's database [2]. The method decides what should be confirmed by minimizing the sum of confirmation cost (C-cost) and information transfer cost (I-cost). The C-cost corresponds to the length of a confirmation subdialogue, and the I-cost to the length of a system response. The C-cost and the I-cost run counter to each other. Although the previous method gives a core idea for controlling a dialogue using the two costs, it defines the C-cost in an unrealistic way, i.e., as the number of the items that the system confirms. It is plausible that the C-cost should depend on the speech recognition rate, and an evaluation of the previous method has not been presented.

In this paper, in line with our previous proposal [2], we present a dialogue control method called the dual-cost method that controls a dialogue so as to minimize the sum of C-cost and I-cost. The C-cost is defined in a more realistic way, namely, as the expected number of content words that are exchanged in a confirmation subdialogue, and it depends on the speech recognition rate. The I-cost is the expected number of content words in a system's response and depends on the content of the system's database. The dual-cost method enables a system to convey necessary information to the user in as short a dialogue as possible depending on the speech recognition rate and the content of its database. The results of dialogue experiments prove that the dual-cost method outperforms conventional methods.

The dual-cost method only allows for "authentic system responses" constructed assuming that only the content of a query that the user has acknowledged is correct. However, we can consider "trial system responses" constructed assuming that unacknowledged content is also correct. When the speech recognition rate is high, trial system responses may have an advantage over authentic ones. We improved the dual-cost method so that the system can make trial responses. The improved method is called the trial dual-cost method. Previous work [1, 4, 5, 6] has dealt with trial system responses in that the models allow for implicit confirmations. Our approach is novel in two respects: we handle trial system responses according to the criterion of minimizing the total dialogue length including the length of a system response and we provide a technique for estimating the total dialogue length depending on the speech recognition and

the content of the system's database. The results of dialogue experiments prove that the trial dual-cost method outperforms the dual-cost method and conventional methods.

2. Spoken Dialogue System

In a spoken dialogue system, the user makes a query about the content of its database and the system conveys the information the user desires. The system can deal with several types of query. For example, consider a weather information system that can handle four types of query: weather categories, temperature, rain probability, and warnings. A dialogue between the system and the user is decomposed into two parts: a confirmation subdialogue and a system response. In a confirmation subdialogue, the system attempts to determine the content of a user query by confirming the content of a recognized query. The system's understanding of a query is represented as a set of triplets comprising an attribute, the value of the attribute, and a propositional constant showing whether the value has already been acknowledged or not. For example, the weather information system needs attributes such as place, date, warning type, and information type. After a confirmation subdialogue, the system makes a response to convey information stored in its database to the user.

The dialogue control here is to choose an optimal system's action at any point of a dialogue so as to minimize the total dialogue length. The system's action is either a confirming action, a soliciting action, or a system response. A confirmation subdialogue is a sequence of confirming or soliciting actions. A confirming action is an action for repeating a confirming question, like "Tokyo?", about the values of attributes until the user acknowledges the confirmation by an affirmative answer like "Yes". A soliciting action is an action for making a soliciting question like "Where?" to obtain the value of an attribute from the user and then performing a confirming action for determining the value. A system response conveys information stored in the system's database to the user. As explained in section 1, there are two types of system response: an authentic response and a trial response.

3. Dual-cost method

The dual-cost method chooses a system's action at each point of dialogue so as to minimize the total dialogue length. The dialogue length is modelled as the sum of the C-cost (the expected number of content words exchanged in a confirmation subdialogue) and the I-cost (the expected number of content words in a system response). The flow of the dual-cost method is as follows.

Step 1 Derive possible types of query under the current system's understanding.

Step 2 Generate possible dialogue plans for each type of query. A dialogue plan is the sequence of system actions to be taken. Possible dialogue plans depend on the type of query.

Step 3 For each dialogue plan, compute the sum of the C-cost and the I-cost. The sum of the costs is the cost of the plan.

Step 4 Generate possible system's actions under the current system's understanding.

Step 5 For each action and for each type of query, select the dialogue plan that contains the action and yields the minimum cost. The minimum cost is the cost of the action for each type of query.

Step 6 Compute the expected cost of each action over the probability distribution of the types of query, and choose the action that yields the minimum expected cost.

Step 7 Execute the chosen action. Wait for user's utterances and update the system's understanding. Go to Step 1.

The dual-cost method generates only dialogue plans including an authentic response in Step 2. Let us explain how the C-cost and the I-cost for a dialogue plan are computed in Step 3. The confirmation cost is accompanied by a confirming action and soliciting action. Consider the expected number of the system-user utterance pairs that occur until a confirming or soliciting action completes. A system-user utterance pair is composed of a system's confirming question or soliciting question and the subsequent user answer. We here make several assumptions. First, the user's answer must be either an affirmative answer like "Yes" or a correcting utterance. Second, the system can correctly recognize a user affirmative answer. Third, the recognition rate of each attribute is given. The recognition rate of an attribute is the probability that the value of the attribute is correctly recognized. The recognition rate for a set of attributes is the probability that the values of all attributes are correctly recognized and is computed as the product of the recognition rate of each attribute.

Consider a confirming action for a set of attributes. Given the recognition rate r for the set of attributes, the expected number of system-user utterance pairs until the confirming action completes is as follows [7]:

$$Pair_c = \sum_{i=1}^{\infty} ir(1-r)^{i-1} = \frac{1}{r} \tag{1}$$

To perform a soliciting action, the system first makes a question like "Where?", and the rest is the same as the process for a confirming action. The expected number of system-user utterance pairs for the soliciting action is as follows:

$$Pair_s = 1 + Pair_c = 1 + \frac{1}{r} \tag{2}$$

Next, consider the number of content words in each pair. We assume that, in each pair, the system confirms all of the attributes. We also assume that, in each pair except for the last one, the user's correcting utterance specifies all the attributes. In the last pair, the user makes an affirmative answer. We approximately estimate the content words in each system's confirming question and user's correcting utterance to be the number of attributes, and count a user's affirmative answer as one. Given the number m of attributes, in each pair except for the last one, the number of content words in the pair is $2m$. In the last pair, the number of content words is $m + 1$. Therefore, the C-cost of a confirming action is

$$2m(Pair_c - 1) + m + 1 = \frac{2m}{r} - m + 1 \tag{3}$$

The C-cost of a soliciting action is

$$2m(Pair_s - 1) + m + 1 = \frac{2m}{r} + m + 1 \tag{4}$$

The C-cost of a dialogue plan is computed using (3) and (4) as the sum of the C-costs of confirming and soliciting actions in the plan. A dialogue plan in the dual-cost method has a single authentic system response. The I-cost of an authentic system response is the expected of content words in the response.

In Step 6, the probability distribution of the types of query under the current system's understanding is approximately computed by the recognition rates of attributes [7].

4. Trial dual-cost method

The trial dual-cost method generates dialogue plans including a trial system response as well as those including an authentic one in Step 2. The C-cost for confirming or soliciting actions and the I-cost for authentic system responses are computed in the same way as in the dual-cost method.

A trial system response is generated assuming that some unacknowledged attribute values are correct. When the assumption is true, the trial response succeeds in that the system successfully conveys the information that the user desires. Otherwise, the trial response fails in that it does not convey that information. The success probability of the trial response is the probability that the unacknowledged attribute values are correct. We assume that the values that the user acknowledged are explicitly specified in a system response and the user is able to notice when the system response fails. We also assume that, when the user notices that the trial system response fails, she repeats the same query until she obtains the desired information.

Now consider a situation where the user is making a query Q of type τ and the system has made a trial system response R_0 assuming the values of some attributes A_0 are correct. Let I_0 be the number of content words in the trial system response R_0. I_0 is the I-cost that must be taken irrespective of the success or failure of R_0. Let q_0 be the success probability of the response R_0. We estimate q_0 as the product of the recognition rates of the assumed attributes A_0. When the response R_0 fails at $1-q_0$, the user repeats the same query Q. After the response R_0 fails, a series of dialogues $D_1, D_2, ...$ for dealing with query Q of type τ takes place until the user obtains the necessary information.

The D_i is a dialogue in which the trial dual-cost method deals with a query of type τ. We suppose that, when the trial dual-cost method carries out a dialogue for dealing with a query of type τ, the average number $\overline{C_\tau}$ of the content words exchanged in a confirmation subdialogue and the average number $\overline{I_\tau}$ of the content words in a system's response are given. In other words, the averages of the C-cost and the I-cost in a dialogue for a query of type τ are given. In the current situation, the average length of each dialogue D_i is counted as $\overline{C_\tau} + \overline{I_\tau}$. We also suppose that, when the trial dual-cost makes a trial response based on unacknowledged attribute values in a dialogue for a query of type τ, the average $\overline{q_\tau}$ of the success probability of the trial response is given. We consider that, in dialogue D_i, a system's response succeeds at probability $\overline{q_\tau}$.

From the above considerations, it is derived that, after the trial response R_0 fails at $1-q_0$, a dialogue with $\overline{C_\tau}+\overline{I_\tau}$ content words repeats $1/\overline{q_\tau}$ times until the user obtains the necessary information. Therefore, the sum of the C-cost and the I-cost of exchanges caused by trial system response R_0 for query Q of type τ is as follows:

$$I_0 + \frac{1-q_0}{\overline{q_\tau}}(\overline{C_\tau} + \overline{I_\tau}) \quad (5)$$

5. Experiments

5.1. Experimental system

To evaluate the proposed methods, we developed a weather information system using SpeechBuilder [3]. The speech recognition, language understanding and generation, and speech synthesis components were constructed by the Japanese version of SpeechBuilder developed in the NTT-MIT collaboration project. The dialogue control component with the system's database was created from scratch and incorporated into the whole architecture by the CGI protocol [3].

The user can make four types of query: weather categories, temperature, rain probability, and warnings. The system's understanding is represented by four attributes: place, date, warning type, and information type. There are 760 places, two dates (today and tomorrow), ten warning types, and four information types (weather, temperature, rain probability and warnings). The database stores the information about the weather categories, the highest and lowest temperature, and six-hour rain probabilities on each date for each place. Also stored is that no warning has been issued anywhere.

5.2. Preliminary experiments

To obtain the speech recognition rate of each attribute, a preliminary dialogue experiment was carried out using three subjects. Each subject made twenty predetermined queries. The types of the queries were randomly chosen. The system was controlled by the dual-cost method, where each attribute has the same recognition rate and the recognition rate from 0.7 to 1.0 was randomly chosen. In each dialogue, the system supposed that attribute-values in the final system's understanding are correct. When attribute values recognized in the course of a dialogue were consistent with the final system's understanding, the values were regarded as correctly recognized. Otherwise, the values were regarded as misrecognized. In this way, the system acquired the recognition rates of each attribute. As a result, we obtained 0.62, 0.96, 0.93 and 0.94 as the recognition rate of the place, the date, the warning-type and the information-type, respectively.

Next, to obtain the constants $\overline{q_\tau}$ and $\overline{C_\tau} + \overline{I_\tau}$ in eq. (5), a second preliminary dialogue experiment was performed using three subjects. Each subject made twenty predetermined queries. The trial dual-cost method controlled the system and utilized the recognition rates obtained in the first preliminary experiment. When the experiment started, constant $\overline{q_\tau}$ was set to 1 and $\overline{C_\tau} + \overline{I_\tau}$ was set to the average number of content words exchanged in a dialogue for each query type in the first preliminary experiments. Whenever a dialogue finished, these constants were updated to the mean of the values that had been observed till then. The constants obtained using sixty dialogues were used in the following evaluation experiment.

5.3. Evaluation experiment

We carried out evaluation experiments in which the trial dual-cost method and the dual-cost method were compared with three conventional methods: the lump-sum method, the piecemeal method, and the no-confirmation method. The lump-sum method confirms as many items as possible at once. The piecemeal method confirms items one by one. The no-confirmation method never makes any confirmation. The dual-cost method, the lump-sum method, and the piecemeal method only deal with authentic system responses. The trial dual-cost method and the no-confirmation method deal with trial system responses.

	Proposed		Conventional		
Query type	TDC	DC	LS	P	NC
Warning	6.0	7.9	13.3	16.9	16.7
Weather	7.4	11.7	14.6	14.8	14.2
Temperature	13.1	16.0	16.3	18.3	16.5
Rain prob.	22.2	24.2	24.4	25.9	32.7

Table 1: The average dialogue length for the trial dual-cost (TDC), the dual-cost (DC), the lump-sum (LS), the piecemeal (P), and the no-confirmation (NC) methods for each query type

In the evaluation experiment, there were fifteen subjects. Each subject made twenty predetermined queries. For each query, the dialogue control method to be applied was determined in advance. Note that each subject did not know the dialogue control method that the system used and the system did not have information about what query the subject was given. The queries were arranged so that each of the five dialogue control methods corresponded to each of the four types of query once.

5.4. Evaluation result

The five dialogue control methods were compared from the perspective of the length of the dialogue. The dialogue length is defined as the number of content words exchanged until the subject obtains the information that she desires. Table 1 shows the average dialogue lengths for five dialogue control methods: the trial dual-cost (TDC), the dual-cost (DC), the lump-sum (LS), the piecemeal (P), and the no-confirmation (NC) methods for each query type. In order to examine whether the average dialogue lengths are significantly different, we utilized the Kruskal-Wallis nonparametric ANOVA with Dunn's nonparametric multiple comparison test ($p < 0.01$). The nonparametric tests were used since we cannot assume that the dialogue lengths are normally distributed.

First, we compared three methods: the dual-cost method, the lump-sum method, and the piecemeal method, which only allow for authentic responses. For a query about warnings and weather categories, the ANOVA indicates a significant difference between the average dialogue lengths, and the multiple comparison test indicates that the average dialogue length for the dual-cost method is significantly shorter than the other two methods. For a query about temperature and rain probability, there are no significant differences.

For a query about warnings, the dual-cost method avoided the confirmation of place. For a query about weather categories, there were cases where the dual-cost method did not confirm the date and conveyed weather categories for both today and tomorrow. However, for a query about temperature and rain probability, the dual-cost method did not avoid a confirmation of place or date since the system response gets too long without it. These results show that, when the dual-cost method can perform a database-dependent dialogue, it outperforms the other two methods. Even when a database-dependent dialogue is not possible, the dual-cost method is not worse than the other two conventional methods.

Next, we compared the five methods, including the trial dual-cost method and the no-confirmation method, which can deal with trial system responses. For a query about warnings and weather categories, the ANOVA indicates a significant difference between the average dialogue lengths, and the multiple comparison test indicates that the average dialogue length for the trial dual-cost method is significantly shorter than for the other four methods. For a query temperature, there is a significant difference between the average dialogue lengths and the average dialogue length for the trial dual-cost method is significantly shorter than for the other methods, excluding the no-confirmation method. There is no significant difference between the trial dual-cost method and the no-confirmation method. For rain probability, there is no significant difference. In cases of a query about rain probability, the system response became so lengthy that the trial dual-cost method seldom took the risk of making trial responses. This result shows that, when the gain of omitting confirmations is judged to be larger than the risk of making a trial response, the trial dual-cost method outperforms the dual-cost methods and the conventional methods, and that, in cases where trial system responses are so risky that they cannot be utilized, the trial dual-cost method is not worse than the other methods.

6. Conclusions

The dual-cost and the trial dual-cost dialogue control methods enable a spoken dialogue system to convey the necessary information to its user in as short a dialogue as possible depending on the speech recognition rate and the content of its database. The trial dual-cost method is an improved version of the dual-cost method, and it can make trial system responses assuming that some unacknowledged attribute values are correct. The dialogue experiment proves that the dual-cost method outperforms the conventional methods and that the trial dual-cost method outperforms the dual-cost method and the conventional ones.

7. Acknowledgements

We thank Dr. Hiroshi Murase and all members of the Dialogue Understanding Research Group for helpful comments.

8. References

[1] Chu-Carroll, J., "MIMIC: an adaptive mixed initiative spoken dialogue system for information queries", Proc. ANLP-2000, 97–104, 2000.

[2] Dohsaka, K., Yasuda, N., Miyazaki, N., Nakano, M. and Aikawa, K., "An efficient dialogue control method under system's limited knowledge", Proc. ICSLP-2000, Vol. 2, 739–742, 2000.

[3] Glass, J. and Weinstein, E., "SpeechBuilder: Facilitating spoken dialogue system development", Proc. Eurospeech-2001, 1335-1338, 2001.

[4] Litman, D. J., Kearns, M. S., Singh, S. and Walker, M. A., "Automatic optimization of dialogue management", Proc. COLING-2000, 2000.

[5] Niimi, Y. and Kobayashi, Y., "Dialog control strategy based on the reliability of speech recognition", Proc. ICSLP-96, 1996.

[6] Roy, N., Pineau, J. and Thrun, S., "Spoken dialogue management using probabilistic reasoning", Proc. ACL-2000, 2000.

[7] Yasuda, N., Dohsaka, K., and Aikawa, K., "Spoken dialogue control based on a turn-minimization criterion depending on the speech recognition accuracy", Proc. 2nd SIGdial Workshop on Discourse and Dialogue, 210–213, 2001.