Information Theoretic Modeling and Development of a Spoken Dialog System

Yasuhisa NIIMI and Yutaka KOBAYASHI
Department of Electronics and Information Science,
Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto, 606 JAPAN
e-mail: niimi@dj.kit.ac.jp

The purpose of this study is to model the dialog control strategy and develop a spoken dialog system based on it. We can view the exchange of information between a spoken dialog system and its user as communication through a noisy channel, because speech recognition errors are inevitable at the current state of the art. In this paper we propose a dialog control strategy to relieve speech recognition errors and investigate by Wizard system factors affecting how comfortably one might converse with such a dialog system.

1. Dialog Control Strategy

In this section we propose a dialog control strategy to relieve speech recognition errors. Let $R(u)$ be a function to measure reliability in recognizing an utterance u. We assume that $R(u)$ is non-negative and reliability is high as $R(u)$ is small. The proposed dialog control strategy, which is illustrated in Fig. 1, can be stated as follows.

1. If $R(u) \leq \theta_1$, the dialog system accepts an utterance u. Let α be the probability that an utterance is accepted, and p the probability that the accepted utterance has been recognized correctly. α and p are generally dependent on θ_1, α being proportional but p inversely proportional to θ_1.

2. It confirms the content of u, if $\theta_1 < R(u) \leq \theta_2$. Let α' be the probability that the confirmation is made, and q be the probability that the confirmed utterance has been recognized correctly.

3. It asks the user to speak again if $R(u) > \theta_2$. The probability for this is $1 - \alpha - \alpha'$.

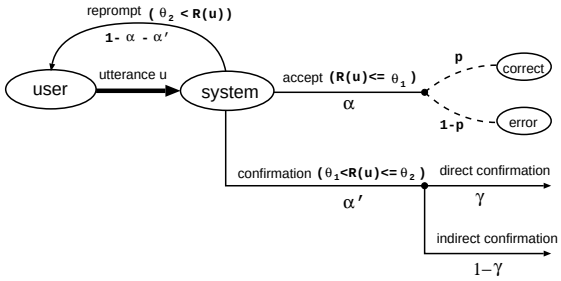

Figure 1: Dialog control strategy

We use two kinds of confirmations; direct confirmation and indirect confirmation. Assume that the dialog system have recognized an utterance as the sentence, “Please tell me

the entrance fee of Kinkakuji temple." The direct confirmation is to ask, "You mean an entrance fee of Kinkakuji temple?", and indirect confirmation is to answer, "You can enter Kinkakuji temple by 500 yen." We select a direct confirmation with the probability γ and an indirect confirmation with the probability $1-\gamma$. We assume the followings for user's response to the confirmation.

1. User's response to the direct confirmation is either "yes" or "no" for simplicity. Following what is stated in (2), an affirmative response occurs with the probability q, and a negative response with the probability $1-q$.

2. The user proceeds to a new utterance without making any comment to the indirect confirmation of correct recognition, but makes some correction to incorrect recognition. Therefore, of user's responses new utterances occur with the probability q and corrections occur with the probability $1-q$.

We also assume the followings for the reaction of the dialog system to user's response denoted by u'. It accepts a response u' if $R(u) \leq \theta_1$, while it rejects u' and asks the user to repeat the first utterance u. Let α be the probability that a response is accepted, and p the probability that the accepted response has been recognized correctly. Then when a response u' to a direct confirmation is accepted and recognized as an affirmative one, the correct information is conveyed to the system with the probability αpq and the incorrect information with the probability $\alpha(1-p)(1-q)$. When the dialog system has accepted u' and recognized as a negative response, it could know it has recognized either u or u' incorrectly. Accordingly the system prompts the user to repeat what he said first. When the user's utterance to an indirect confirmation is accepted and recognized as a new utterance (its probability being αpq), we consider the turns taken for the indirect confirmation and user's new utterance are not spent to convey the information of the first utterance u.

The purpose of modeling the dialog control strategy is to estimate two quantities P_{ac} and N; P_{ac} is the probability that information included in user's utterance is conveyed to the system correctly, and N is the average number of turns taken between the user and the system until termination of subdialog on user's first utterance. Using the method stated in [1], we have the following formulae for P_{ac} and N;

$$
\begin{aligned}
P_{ac} &= \frac{p\{1+\alpha'[1+\gamma(q-1)]\}}{1+\alpha'[1+\gamma(\beta-1)]} \\
N &= \frac{1}{\alpha\{1+\alpha'[1+\gamma(\beta-1)]\}} \times \{2-\alpha+\alpha'[(4-\alpha\beta)\gamma+(4-\alpha-2\alpha pq)(1-\gamma)-2]\}
\end{aligned}
$$

where $\beta = 1+2pq-p-q$.

Here we consider the simplest case of the dialog control strategy mentioned above, that is, the case where $\alpha' = 0$. Letting $P_{ac}^{(0)}$ and $N^{(0)}$ be the two parameters for this case, we have

$$P_{ac}^{(0)} = p \quad \text{and} \quad N^{(0)} = 2/\alpha - 1.$$

By simple calculation we have the following inequalities,

$$P_{ac} > P_{ac}^{(0)} \qquad \text{if } q > 1/2$$

and $N < N^{(0)}$ if $\gamma(p+q+(\alpha-2)pq) < \alpha pq$. When these conditions are satisfied, we can increase P_{ac} and reduce N by making confirmations.

Now we consider how to estimate the parameters α, α', p and q. Let A denote the acoustic data stream of an utterance, and W denote a string of words. We can adopt the conditional probability $P(W/A)$ of W given A as a recognition score. The recognized string of words is such a string that maximizes $P(W/A)$ under the given linguistic constraint. By Bayes' theorem,

$$P(W/A) = P(A/W)P(W)/P(A).$$

The quantity $P(A/W)P(W)$, which is used as a conventional criterion in speech recognition, is computed by using the hidden Markov model and the language model. $P(A)$ is approximated by $\max(P(X)P(A/X))$ where X is a string of phonemes. $R(u)$ is computed by using $P(W/A)$. Using this scheme to compute $R(u)$'s for many training utterances, we can create a distribution for $R(u)$. Selecting a threshold value θ_1, we can estimate α as the area of the portion of the distribution in which $R(u) \leq \theta_1$ is satisfied, and α' as the area of the portion of the distribution in which $\theta_1 < R(u) \leq \theta_2$ is satisfied. p and q are also estimated in the similar way by using separate distributions created from correct recognitions and incorrect recognitions.

Using a continuous speech recognition system, SUSKIT-II [2], we estimated $R(u)$, and thereby the four parameters α, α', p and q. Then we evaluated the proposed dialog control strategy quantitatively by computing P_{ac} and N. It has turned out that while the increase of P_{ac} by the direct confirmation is negligible, the reduction of N is remarkable by the indirect confirmation.

2. Overview of Spoken-Dialog Collection Experiments

2.1 WOZ dialog system

In order to realize a practical human-machine spoken-dialog system, each aspect of human dialog must be well studied and modeled in the system. The authors have pointed out the fact that dialog control strategies affect user's attitude and utterances to the system, and proposed a hierarchical dialog control strategy. In this section, among many factors, comfortableness in the spoken dialog was studied with respect to system's recognition errors and response strategies.

We carried out a preliminary wizard dialog experiment. A user and the person who play the role of the system sit in two separate rooms. They communicate exclusively by speech using head sets. The system's speech is transformed by a commercial voice-changing microphone so that the user would imagine a computer is generating responses. In front of the user there is a bitmap display, on which a piece of picture and a short message are displayed in order to inform him the system's internal state. The three system states correspond to waiting for user's utterance, responding and retrieving information from the database, respectively. The user cannot speak during the periods of the latter two states. The system actor changes the picture and message by simply hitting a key.

2.2 Task domain

The user was asked to build a one-day sight-seeing plan in Kyoto city by asking to the system about popular spots, restaurants, hotels and transportations. Before starting conversation, the user is given a rough map of the city and a plan sheet to fill in, in order to clarify the intention of the dialog.

Goal-oriented dialogs in a limited task domain would be keeping their position for a while as a major target of studies on spoken dialog. Such dialogs can be classified with respect to the goal of conversation:

1. the user extracts some information from the system, and

2. the system extracts some information from the user.

Data retrieval tasks drop in the first category, while conference registration and ticket reservation tasks in the second. An actual conversation would be a mixture of both of these two categories. The user would expect the system to build some task-dependent representation of the user's intentions exchanged during the dialog.

67 WOZ-dialog from 43 subjects (students) were recorded, transcribed and analyzed.

2.3 Controlling system utterances

There are many factors which affect the efficiency of information transfer during spoken dialog and its comfortableness. In addition to the contents of the text, the prosody of the partner's utterances imposes a lot of influence.

Under the human-machine spoken-dialog environment which is not free from recognition error, a friendly response strategy should be developed. In the experiment, we focused on the effects of the system's recognition rate and the confirmation strategies over the comfortableness of the dialog.

3. Comfortableness of Dialog

The comfortableness of the dialog is not less important than the efficiency of information transfer in a dialog.

Since the perplexity of the grammar is a popular measure of difficulty in speech recognition, it is clear that the smaller number of possibilities of the user's next topics would decrease the recognition errors. This can be realized by writing an easier grammar as well as an appropriate use of pragmatics. However, in a spoken dialog system, one can make a separate effort by asking simpler questions, dividing the selection list, and so forth. In other words, the dialog control of choice selection level should take care of facilitating the problem that the speech recognition system must solve, fundamentally independent of the grammar and higher intelligent processings.

In case of high error rate, the system needs frequent confirmation dialogs, suspecting its own recognition results. Otherwise, it would be more convenient to have a confirmation dialog at the end of each dialog block. When and how they should be done are considered as a problem of communication protocols. The intensive study on efficient protocols which take into account the characteristics of speech recognition errors. Especially in the corrective dialog, the communication level control must avoid any additional errors.

In this section we applied a multi-variate analysis to the correlations among the user's impression of comfortableness and several measurable features of the dialog.

3.1 Regression equation of comfortableness

The 51 sets of following seven values were measured for each dialog:

Table 1: Normalized regression coefficients

variables	regression coeffs.	t-value
x_1	b_1= 0.009	0.04
x_2	b_2=-0.259	**2.10**
x_3	b_3= 0.002	0.01
x_4	b_4= 0.189	0.84
x_5	b_5= 0.123	0.90
x_6	b_6=-0.707	**4.45**
Variance (F)	**7.29** $> F(6,44;0.01) = 3.243$	
Contribution (R^2)	0.50	

x_1 : the length of the system's utterances (sec)
x_2 : the total length of the dialog (sec)
x_3 : the efficiency of the conversation (items/sec)
x_4 : the rate of indirect confirmation responses (%)
x_5 : the rate of direct confirmation responses (%)
x_6 : the system's recognition error rate (%)
y : the user's rating of comfortableness (%)

and estimated the regression coefficients of the target variable (comfortableness) with respect to the other six variables. x_3 is the average information transfer rate per a second. y was obtained from the enquête filled out after the recording of each dialog.

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_6 x_6$$

The normalized regression coefficients b_i are shown in Table .

3.2 Validity of the equation

Using the regression equation, the regression estimate Y_λ is calculated for each sample value y_λ of the target variable. For the sake of testing, the validity estimate (t-value) of each variables are tabulated in the last column of Table . The variance ratio (F) and the contribution (R^2) of the estimates to the variance are calculated by:

$$F = \frac{RV/p}{EV/N-p-1}, \qquad R^2 = RV/a_{yy}.$$

where a_{yy} is the total variance of y_λ, RV is the variance of Y_λ, and EV is that of the residual ($\epsilon = y_\lambda - Y_\lambda$).

Variables x_2 and x_6 were accepted as valid at confidence level 95%, while x_1 and x_3 had little effect on the comfortableness. As naturally understood, the system's recognition error has a strong negative influence. One can say the frequencies of both confirmation responses, x_4 and x_5, contribute to the comfortableness slightly.

4. Summary

This paper has reported a dialogue control strategy to relieve errors in speech recognition, and analyzed them mathematically. The analysis has proven that the direct confirmation can increase the probability that information included in user's utterances is conveyed to

the system correctly, and the indirect confirmation can reduce the average number of turns exchanged between the user and the system.

Through the multi-variate analysis of several control factors of spoken-dialog, we found confirmations and rephrasing responses do not decrease the comfortableness of conversations.

Bibliography

[1] Niimi,Y. and Kobayashi,Y., "Modeling Dialogue Control Strategies to Relieve Speech Recognition Errors," Proc. of EUROSPEECH'95, pp.1177–1180 (1995).

[2] Kobayashi,Y., Omote,M., Endo,H. and Niimi,Y., "SUSKIT-II — A speech understanding system based on robust phone spotting," IEICE Trans., Vol. E-74, No. 7, pp.1863–1869 (1991).

Keywords: dialog control strategy, spoken dialog system, dialog model, confirmation, WOZ experiment, comfortableness