A Model of Temporally Changing User Behaviors in a Deployed Spoken Dialogue System

Kazunori Komatani, Tatsuya Kawahara, and Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University,
Yoshida-Hommachi, Sakyo, Kyoto 606-8501, Japan
komatani@i.kyoto-u.ac.jp

Abstract. User behaviors on a system vary not only among individuals but also within the same user when he/she gains experience on the system. We empirically investigated how individual users changed their behaviors on the basis of long-term data, which were collected by our telephone-based spoken dialogue system deployed for the open public over 34 months. The system was repeatedly used by citizens, who were each identified by their phone numbers. We conducted an experiment by using these data and showed that prediction accuracy of utterance-understanding errors improved when the temporal change was taken into consideration. This result showed that modeling temporally changing user behaviors was helpful in improving the performance of spoken dialogue systems.

Keywords: Spoken dialogue system, temporal change, real user behavior, habituation, barge-in, deployed system.

1 Introduction

User behaviors are an important factor that should be considered when designing a spoken dialogue system and improving its performance. We empirically investigated how individual users became skilled in using the system. We used long-term data collected by our telephone-based spoken dialogue system [1] used by the general public. We assumed each individual is identified by their telephone number. We analyzed several user behaviors per individual including barge-in rate. A barge-in is a situation in which a user starts speaking during a system prompt and is a characteristic feature of spoken dialogue systems. The barge-in rate reveals in what manner a user uses the system to complete a task.

Our study is characterized by capturing temporal changes of individual users as they acquire experience in using the system. Walker et al. developed a user model that applies to general users and constructed a spoken dialogue system adapted to them [2]. We constructed an individual user model in spoken dialogue systems, based on the classification of Jameson and Wittig [3], after investigating real user behaviors. Another characteristic of our study is that we exploit the barge-in rate as a new profile at the dialogue level. Some studies have used

G.-J. Houben et al. (Eds.): UMAP 2009, LNCS 5535, pp. 409–414, 2009.

dialogue-level features to detect ASR errors [4,5,6]. With respect to the barge-in in spoken dialogue systems, Ström and Seneff discussed how to manage barge-in detection errors [7], and Rose and Kim showed an experimental study how barge-in detection errors affected user utterances [8]. However, the locutionary-act-level phenomenon, barge-in, has not been exploited to detect ASR errors.

We used the barge-in rate to predict errors for "barge-in utterances," where a user barges in with an utterance during the system prompt. We set a window when calculating the barge-in rates to reflect temporally changing user behaviors by discarding their old histories. We show how the prediction accuracy improved when we took the temporal change into consideration.

2 Target Data from Deployed Spoken Dialogue System

We developed the Kyoto City Bus Information System [1] that received user utterances and provided information all by voice. The system locates a bus that a user wants to catch and tells him/her how long it will be before the bus arrives. The system was open to the public and was accessible by telephone, including cellular phones. It operated on a product of Nuance Communications, Inc.

We used data collected by the system between May 2002 and February 2005. The data contained 7,988 valid calls from 671 users. Callers' phone numbers were recorded for 5,927 of the 7,988 calls. We analyzed behaviors of individual users based on these phone numbers. Each utterance was transcribed, and then the language understanding result, whether correct or not, was given manually. We assumed that a language understanding result for an utterance was correct if all content words in its transcription were correctly included in the result. It was regarded as an error if any content words were not correctly recognized in automatic speech recognition (ASR). As with the language understanding results, a task success was also determined manually.

We counted how many times each user called the system. The result is listed in Table 1. Note that the numbers of tasks are not equal to the number of calls multiplied by the number of users because some users completed several tasks during a single call or hung up before completing tasks. We can see a tendency that task success rates were higher as the number of calls per user increased. The number of users who used the system only once during this period was 306, representing 45.6% of total users. Twelve users, meanwhile, called the system over 50 times. All of the twelve phone numbers were those of mobile phones, which are generally not shared, so we can expect that each number corresponds to individuals.

3 Analyzing Temporal Transitions of User Behaviors

Users are expected to change their behaviors, such as how often they barge-in, until they get sufficiently accustomed to the system. We analyzed temporal transitions of user behaviors including barge-in rates. Results for ASR accuracy and task success rate and their relations can be found in [9]. The barge-in rate

Table 1. Number of users per number of calls and their task success rates

# of calls	# of users	Task success rate (%) (#Succeeded/#Tasks)
1	306	76.4 (191/250)
2	130	76.1 (169/222)
3	69	72.1 (124/172)
4	31	71.4 (85/119)
5-9	61	77.0 (285/370)
10-19	39	84.1 (419/498)
20-29	13	92.3 (251/272)
30-39	8	92.7 (229/247)
40-49	2	88.9 (72/ 81)
50-99	6	88.9 (408/459)
100-199	1	94.5 (137/145)
200-299	1	97.1 (298/307)
300-399	1	90.8 (314/346)
400-499	2	95.7 (900/940)
500-599	1	94.2 (491/521)
Total	671	88.4 (4347/4949)

Table 2. Temporal transitions of barge-in rates for frequent users

User ID	$f(1)$	Δ	x_I	MSE
#1	.11	0	-	2.3E-4
#2	.19	0	-	1.9E-3
#3	.60	.60	> 1	6.4E-4
#4	.17	0	-	7.2E-4
#5	.74	.74	.58	4.6E-4
#6	.10	.06	< 0	1.1E-4
#7	.04	.04	.06	1.6E-4
#8	.71	0	-	1.0E-3
#9	.49	.47	.62	4.6E-4
#10	.10	.10	.29	1.3E-4
#11	.15	.04	.13	9.8E-4
#12	.23	0	-	2.6E-3
Average	.30	.17	-	-
Stdev	.24	.26	-	-

MSEs: mean square errors

was defined as the ratio of the number of utterances in which a user barges-in on system prompts and the number of total utterances performed by the user.

Temporal transitions of the barge-in rates for users #1 and #5 are shown in Figure 1 as examples. As a temporal axis, we calculated the ratios using the number of utterances up to a certain point and the number of total utterances by the user, and plotted them on the x-axis. Therefore, $0 < x \leq 1$. Average barge-in rates per user to a certain time x were plotted on the y-axis. The examples show that barge-in rate of user #1 was nearly static, whereas the barge-in rate of user #5 increased as they became used to the system. As highlighted by these examples, variations in barge-in rates depended on individual users.

We then approximated the plotted values by using the following function: $f(x) = c - a \cdot \exp(-bx)$. These parameters were calculated by using the least squares method. We assumed $a \geq 0$. To describe rough shapes of the approximation functions, three values were calculated such as $f(1)$, Δ, and x_I. Here, $f(1)$ represents an average of each measure in this period. Δ was defined as $f(1) - f(0)$, which represents the change of each measure for the user in this period. We calculated x_I as $\{x | \frac{df(x)}{dx} = 0.1\}$, which means that the change of $f(x)$ converges near x_I. Note that x_I is not defined when Δ is zero because there is no change in $f(x)$. Table 2 summarizes temporal transitions of the 12 users who used the system more than 50 times. The table shows that barge-in rates of some users, such as users #3, #5, and #9, increased steeply, whereas the rates of the other users did not change very much. Standard deviations of the averages ($f(1)$) and the amount of change (Δ) were rather large, which showed the diversity of the user behavior.

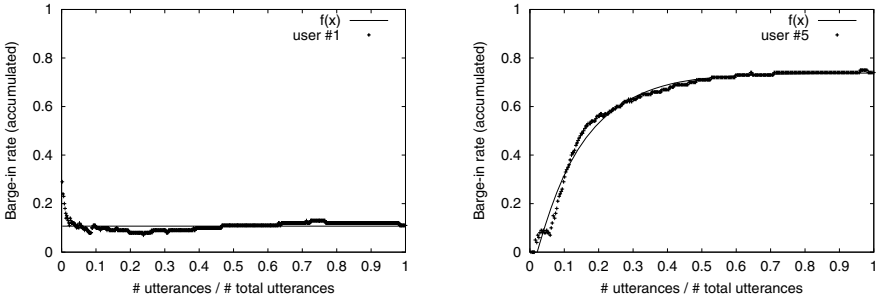

Fig. 1. Temporal transition of barge-in rate for users #1 and #5

4 Predicting Errors by Using Temporally-Changing Barge-in Rate

We conducted an experiment to verify whether the model of temporal change of user behaviors was helpful for improving performance of a spoken dialogue system. We considered the model when predicting utterance-understanding errors of barge-in utterances on the basis of each user's barge-in rate.

Barge-in utterances are prone to containing more ASR errors than those without barge-ins. The barge-in utterances amounted to 26.8% (7,940/29,580) of all utterances, and about half of those contained utterance-understanding errors caused by ASR errors [11]. These were caused by background noise, disfluencies in user utterances, or the user's unfamiliarity with the system. ASR errors often occur in fragments of utterances, especially when novices use the system [10] and cause utterance-understanding errors as a result. An example is when users were not accustomed to the timing when to speak and stopped their utterances when they noticed the system prompt continued. Disfluencies are another reason as Rose and Kim reported that more disfluencies appeared when users barged in compared to when users waited until the prompt ended [8].

4.1 Predicting Errors on the Basis of Barge-in Rate

We had confirmed the relationship between the average barge-in rate per user and the corresponding utterance-understanding accuracy of barge-in utterances [11]. For users whose barge-in rates were high, that is, they frequently barged-in, the utterance-understanding accuracy of barged-in utterances was high. This suggests that the barge-ins were done intentionally. On the other hand, for users whose barge-in rates were low, their utterance-understanding accuracies of such utterances were low, too. This suggests that the barge-ins might be unintentional.

To predict utterance-understanding errors from the barge-in rate, we used a logistic regression model. Denoting a probability that an utterance-understanding result of a barge-in utterance is correct as P, the regression function is written as:

$$P = \frac{1}{1 + \exp(-(a_1 x_1 + a_2 x_2 + b))}.$$

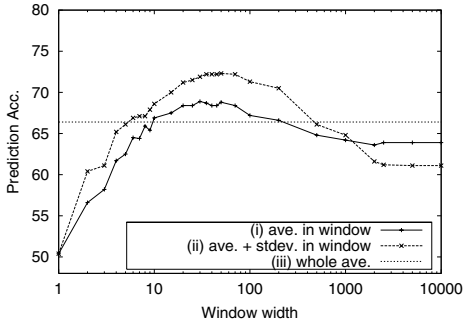

Fig. 2. Prediction accuracy when window width varies

Table 3. Best prediction accuracy and corresponding window width

(i)	(ii)	(iii)	Maj.
68.9%	72.3%	66.4%	50.4%
(w=30)	(w=50)	(-)	(-)

Figures in () are window width.

The independent variables are x_1 and x_2, which represent the average and standard deviation of barge-in rates, respectively. The dependent variable is a binary value indicating whether the utterance-understanding result is correct or not. The coefficients a_1, a_2, and b are obtained after fitting by using training data.

In order to take the temporal transition of user behaviors into consideration, we set a window for calculating barge-in rates at each point of the dialogue. That is, barge-in rates are calculated by using N utterances before the current one. We call this N the window width. When too wide a window is used, the average barge-in rate does not reflect the temporal change of user behaviors. When a window is too narrow, the average barge-in rate is not reliable as a user profile. Standard deviations of barge-in rates are also calculated within the window. Small standard deviations mean that the barge-in rate has already converged, and accordingly its average can be used as a reliable profile.

4.2 Experimental Verification

We set the following three experimental conditions: (i) only used average barge-in rates (x_1), and (ii) used both averages and standard deviations of barge-in rates (x_1, x_2) within each window width. Condition (iii) used the average barge-in rate per individual calculated by using all utterances and did not take into consideration temporal transition of user behaviors.

We calculated the prediction accuracy by using all 7,940 barge-in utterances. The fitting and prediction processes were performed by a 10-fold cross validation. When a window width exceeded the number of all utterances by the user, barge-in rates were calculated by using the all utterances. Figure 2 shows the prediction accuracies when the window width varies. Accuracies and window widths when the best performance was obtained are listed in Table 3. "Maj." in this table means the majority baseline, that is, when all utterances were classified to either binary value.

Prediction accuracies for (i) and (ii) with appropriate window widths were better than (iii) (i.e., when the average of all utterances were used), as shown

in Figure 2. The use of the standard deviation, as shown in Condition (ii), also improved the prediction accuracy. The window discarded the users' old histories and thus reflected temporal transitions of their behaviors. Consideration of temporal transitions improved the performance of (i) and (ii), because the barge-in rates were not constant but varied as the users got accustomed to the system, as shown in Figure 1. Figure 2 also shows that prediction accuracy leveled off for window widths larger than around 30. This means that several dozens of utterances at least need to be used to calculate the average barge-in rate. Each call contained 2-6 utterances, so reliable histories are formed when a person used the system more than about 10 times.

As a conclusion, the temporal model was effective and should be considered for improving the system performance. We will further investigate users' actual intentions when they barge in a system prompt. Integration with other measures such as ASR confidences are also included in our future work.

References

1. Komatani, K., Ueno, S., Kawahara, T., Okuno, H.G.: User modeling in spoken dialogue systems to generate flexible guidance. User Modeling and User-Adapted Interaction 15(1), 169–183 (2005)
2. Walker, M., Langkilde, I., Wright, J., Gorin, A., Litman, D.: Learning to predict problematic situations in a spoken dialogue system: Experiments with how may I help you? In: Proc. NAACL, pp. 210–217 (2000)
3. Jameson, A., Wittig, F.: Leveraging data about users in general in the learning of individual user models. In: Proc. IJCAI 2001 (2001)
4. Litman, D.J., Walker, M.A., Kearns, M.S.: Automatic detection of poor speech recognition at the dialogue level. In: Proc. ACL, pp. 309–316 (1999)
5. Gabsdil, M., Lemon, O.: Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems. In: Proc. ACL, pp. 343–350 (2004)
6. Bohus, D., Rudnicky, A.: A "k hypotheses + other" belief updating model. In: Proc. AAAI Workshop on Statistical and Empirical Approaches to Spoken Dialogue Systems (2006)
7. Ström, N., Seneff, S.: Intelligent barge-in in conversational systems. In: Proc. ICSLP, pp. 652–655 (2000)
8. Rose, R., Kim, H.: A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems. In: Proc. of ASRU, pp. 198–203 (2003)
9. Komatani, K., Kawahara, T., Okuno, H.G.: Analyzing temporal transition of real user's behaviors in a spoken dialogue system. In: Proc. INTERSPEECH, pp. 142–145 (2007)
10. Raux, A., Bohus, D., Langner, B., Black, A., Eskenazi, M.: Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In: Proc. INTERSPEECH, pp. 65–68 (2006)
11. Komatani, K., Kawahara, T., Okuno, H.G.: Predicting asr errors by exploiting barge-in rate of individual users for spoken dialogue systems. In: Proc. INTERSPEECH, pp. 183–186 (2008)