Prediction of Prospective User Engagement with Intelligent Assistants

Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano
Yahoo Japan Corporation
9-7-1 Akasaka, Minato-ku, Tokyo 107-6211, Japan
{shsano, nkaji, msassano}@yahoo-corp.jp

Abstract

Intelligent assistants on mobile devices, such as Siri, have recently gained considerable attention as novel applications of dialogue technologies. A tremendous amount of real users of intelligent assistants provide us with an opportunity to explore a novel task of predicting whether users will continually use their intelligent assistants in the future. We developed prediction models of prospective user engagement by using large-scale user logs obtained from a commercial intelligent assistant. Experiments demonstrated that our models can predict prospective user engagement reasonably well, and outperforms a strong baseline that makes prediction based past utterance frequency.

1 Introduction

Intelligent assistants on mobile devices, such as Siri,[1] have recently gained considerable attention as novel applications of dialogue technologies (Jiang et al., 2015). They receive instructions from users via voice control to execute a wide range of tasks (*e.g.*, searching the Web, setting alarms, making phone calls, and so on). Some are able to even chat or play games with users (Kobayashi et al., 2015).

Intelligent assistants possess a unique characteristic as an object of dialogue study. Popular intelligent assistants have thousands or even millions of real users, thanks to the prevalence of mobile devices. Some of those users continually use intelligent assistants for a long period of time, while others stop using them after a few trials. Such user behaviors are rarely observed in conventional experimental environments, where dialogue systems have only a small number of experimental participants who almost always continue to use the systems for the whole duration of the experiment.

This paper explores a novel task of predicting whether a user will continue to use intelligent assistants in the future (This task is referred to as *prospective user engagement prediction* and its definition is given in Section 3). We attempt to develop such a prediction model, which would contribute to enhancing intelligent assistants in many ways. For example, if users who are likely to stop using systems can be identified, intelligent assistants can take actions to gain or maintain their interest (*e.g.*, by sending push notifications).

This task is related to, but is significantly different from, user engagement detection, which has been extensively explored in prior dialogue studies (Wang and Hirschberg, 2011; Forbes-Riley et al., 2012; Forbes-Riley and Litman, 2013; Oertel and Salvi, 2013). The prior studies attempt to predict how strongly users are currently engaged in dialogues with systems. On the other hand, the goal of this study is to predict how strongly users will be engaged with intelligent assistants in the future. The largest difference lies in whether the prediction target is user engagement at present or in the future. Also, our definition of engagement is slightly different from the prior ones. In this study, engagement is considered as a sentiment as to whether users like intelligent assistants and feel like they want to use them continually.

To develop and evaluate models of prospective user engagement prediction, we exploit large-scale user logs obtained from a commercial intelligent assistant. Since monitoring users' long-term behaviors is considered crucial for precise prediction of their prospective engagement, we tailor various features by extracting usage patterns from a long history of user dialogues. The resulting features are contrastive to those previously used for user

[1] http://www.apple.com/ios/siri

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1203–1212,
Berlin, Germany, August 7-12, 2016.

engagement detection, in which features are basically extracted from a single user utterance.

Experimental results demonstrated that our models are able to predict prospective user engagement reasonably well and are overwhelmingly better than a strong baseline that makes predictions based on past utterance frequency. We also discuss the trade-off between prediction accuracy and instancy. Specifically, we investigate how the prediction performance improves as we wait for more user dialogues to be collected.

2 Yahoo! Voice Assist

This section summarizes *Yahoo! Voice Assist*[2] (hereafter Voice Assist), a commercial intelligent assistant that is investigated in this study. Although our investigation exclusively focused on this system, we will discuss how our findings can be generalized to other intelligent assistants in Section 5.5

Table 1 illustrates example dialogues of Voice Assist users.[3] As illustrated, Voice Assist offers a variety of functions to mobile users. They are largely categorized into two types:

device operation Voice Assist allows users to operate mobile devices through dialogue. This includes setting alarms, making phone calls, searching the Web, launching an app, and so on (*e.g.*, **V1**, **V3**, **V4**, and **V5**).

chat Voice Assist can give greetings to, have conversations with, and play games with users (*e.g.*, **V2** and **V6**). In contrast to device operations for accomplishing certain tasks, these functions are offered for fun or for facilitating smooth communication.

Voice Assist currently supports 66 functions (including setting alarms, the word chain game, etc.) and they can further be classified into fine-grained types, although a detailed description of them is beyond the scope of this paper.

Voice Assist users can register personal profile such as their home address and birthday, with which the system makes personalized responses. For example, the home address are used to estimate users' location when launching weather forecast apps (*i.e.*, response **V3**), while knowing birthdays allows Voice Assist to send greeting messages to users on their birthdays.

U1	Wake me up at 8 o'clock tomorrow.
V1	OK. Set the alarm to 8 am. (show the timer configuration)
U2	Good morning.
V2	Good morning. Thank you for talking to me.
U3	Check today's weather.
V3	Tokyo's weather will be fine today. (launch weather forecast app)
U4	From Tokyo to Kyoto.
V4	A rapid express is available at 9:30. (launch transit app to show timetable)
U5	What time will it arrive?
V5	It will arrive at Kyoto at 11:50. (show the timetable again)
U6	Let's play the word chain game.
V6	OK. Apple...

Table 1: Example dialogues of Voice Assist users. **U** and **V** indicate the user and Voice Assist, respectively. The notes in parentheses represent actions that Voice Assist takes after the responses.

3 Prospective User Engagement Prediction

This section specifies the task of prospective user engagement prediction. We first explain the user log data used in our experiments. We then describe two kinds of task settings.

3.1 User log data

We conducted an empirical study in which we examined Voice Assist user logs. We randomly sampled 348,295 users who used the system at least once from March 2015 to June 2015 (16 weeks) and extracted all their dialogue histories during that period. The log data included 7,472,915 utterances in total.

Table 2 illustrates examples of user logs. We used the following seven attributes: user ID, nickname, birthday, time stamp, user utterance, system response, and response type. Because it is not mandatory to register the personal profiles (including nicknames, birthdays, etc.), they are sometimes missing, as indicated by N/A in the table. The response type represents the 66 functions supported by Voice Assist. The time stamps were used to segment utterances into sessions, as rep-

[2] `http://v-assist.yahoo.co.jp` (in Japanese)

[3] Because Voice Assist supports only Japanese, all utterances are made in Japanese. In this paper, we present English translations rather than the original Japanese to facilitate non-Japanese readers' understanding.

ID	Nickname	Birthday	Time Stamp	Use Utterance	System Response	Type
A	John	2000-1-1	2015-3-1 23:50	Wake me up at 8 am tomorrow.	OK. Set the alarm to 8 am.	ALARM
			2015-3-2 08:10	Good morning.	Good morning.	CHAT
			2015-3-2 08:13	Check today's weather.	Tokyo's weather will be fine today.	WEATHER
B	N/A	2002-1-1	2015-3-1 08:00	From Tokyo to Kyoto.	A rapid express is available at 9:30.	TRANSIT
			2015-3-1 08:01	What time will it arrive?	It will arrive at Kyoto at 11:50.	TRANSIT
			2015-3-5 19:10	Let's play the word chain game.	OK. Apple...	WORD CHAIN

Table 2: User log examples. The dashed line represents the session boundary.

resented by dashed lines in the table. We follow (Jiang et al., 2015) to define sessions as utterance sequences in which the interval of two adjacent utterances does not exceed 30 minutes.

3.2 Task definition

We propose two types of prospective user engagement prediction tasks. In both tasks, we collect user dialogues from the first eight weeks of the user logs (referred to as *observation period*. We will discuss on length of observation period in Section 5.4), and then use those past dialogues to predict whether users are engaged with the intelligent assistant in the last eight weeks of the log data (referred to as *prediction period*).[4] We specifically explored two prediction tasks as follows.

Dropout prediction The first task is to predict whether a given user will not at all use the system in the prediction period. This task is referred to as *dropout prediction* and is formulated as a binary classification problem. The model of dropout prediction would allow intelligent assistants to take proactive actions against users who are likely to stop using the system. There are 71,330 dropout users, who does not at all use the system in the prediction period, among 275,630 in our data set.

Engagement level prediction The second task aims at predicting how frequently the system will be used in the prediction period by a given user. Because there are outliers, or heavy users, who use the system extremely frequently (one user used the system as many as 1,099 times in the eight weeks), we do not attempt to directly predict the number of utterances or sessions. Instead, we define engagement levels as detailed below, and aim at predicting those values.

The engagement levels are defined as follows. First, users are sorted in the ascending order of the number of sessions they made in the prediction period. We then split users into four equally-sized groups. The engagement levels of users in the four groups are defined as 1, 2, 3, and 4, respectively (Table 3). Note that a larger value of the engagement level means that the users are more engaged with the intelligent assistants. This task is referred to as *engagement level prediction* and is formulated as a regression problem.

Level	# of sessions	# of users
1	0	71,330
2	1–3	66,626
3	4–13	69,551
4	14–	68,123

Table 3: User distribution over the four engagement levels. The second column represents intervals of the number of sessions corresponding to the four levels.

The engagement level prediction has different applications from the dropout prediction. For example, it would allow us to detect in advance that a user's engagement level will change from four to three in the near future. It is beyond the scope of dropout prediction task to foresee such a change.

4 Features

The dropout prediction is performed using linear support vector machine (SVM) (Fan et al., 2008), while the engagement level prediction is performed using support vector regression (SVR) (Smola and Schölkopf, 2004) on the same feature set. Here, we divide the features into four categories by their function: utterance frequency features, response frequency features, time interval features, and user profile features. Table 4 lists these features.

4.1 Utterance frequency features

Here, we describe the features related to utterance frequency. These features attempt to capture our

[4] We removed users from the log data if the number of sessions was only once in the observation period, because such data lack a sufficient amount of dialogue histories for making a reliable prediction.

#Features	Name	Definition
1	**Utterance**	The number of utterances
7	**Utterance**w**Weeks**	The number of utterances in recent w weeks
1	**LongUtterance**	The number of lengthy utterances
1	**UrgedUtterance**	The number of utterances made in response to push notifications
1	**Restatement**	The number of restatement utterances
100	**UtteranceTopic**i	The number of utterances including words in the i-th cluster
1	**Session**	The number of sessions
7	**Session**w**Weeks**	The number of sessions in recent w weeks
7	**SessionByDay**	The number of sessions during each day of the week
66	**Response**(t)	The number of responses with response type t
66	**FirstResponse**(t)	**Response**(t) computed by using only the first responses in sessions
1	**LongResponse**	The number of lengthy responses
1	**ErrorMessage**	The number of error messages
1	**MaxInterval**	Max days between adjacent utterances
1	**MinInterval**	Min days between adjacent utterances
1	**AvgInterval**	Average days between adjacent utterances
1	**InactivePeriod**	Days from the last utterance date
66	**InactivePeriod**(t)	**InactivePeriod** computed for each type of the last response
1	**Nickname**	Whether or not a user has provided nickname information
1	**Birthday**	Whether or not a user has provided birthday information
6	**Age**	User's age category

Table 4: List of features. The utterance frequency features, response frequency features, and time interval features are all scaled.

intuition that users who frequently use intelligent assistants are likely to be engaged with them.

Utterance The number of utterances in the observation period. For scaling purposes, the value of this feature is set to $\log_{10}(x+1)$, where x is the number of utterances. The same scaling is performed on all features but user profile features.

Utterance$w$**Weeks** The number of utterances in the last w $(1 \leq w < 8)$ weeks of the observation period.

LongUtterance The number of lengthy utterances (more than 20 characters long). Jiang et al. (2015) pointed out that long utterances are prone to cause ASR errors. Since ASR errors are a factor that decreases user engagement, users who are prone to make long utterances are likely to be disengaged.

UrgedUtterance The number of utterances made in response to push notifications sent from the system. We expect that engaged users tend to react to push notifications.

Restatement The number of restatements made by users. Jiang et al. (2015) found that users tend to repeat previous utterances in case of ASR errors. An utterance is regarded as a restatement of the previous one if their normalized edit distance (Li and Liu, 2007) is below 0.5.

UtteranceTopic$i$ The number of utterances including a keyword belonging to i-th word cluster. To induce the word clusters, 100-dimensional word embeddings are first learned from the log data using WORD2VEC (Mikolov et al., 2013)[5], and then K-means clustering (K=100) is performed (MacQueen, 1967). All options of WORD2VEC are set to the default values. These features aim at capturing topics on utterances or speech acts. Table 5 illustrates example words in the clusters. For example, utterances including words in the cluster ID 36 and 63 are considered to be greeting acts and sports-related conversations, respectively.

[5] https://code.google.com/archive/p/word2vec

Cluster ID	Example words
14 (Weather)	pollen, typhoon, temperature
23 (Curse)	die, stupid, shit, shurrup, dorf
36 (Greeting)	thanks, good morning, hello
48 (Sentiment)	funny, cute, good, awesome
63 (Sports)	World cup, Nishikori, Yankees

Table 5: Example words in the clusters. Cluster names (presented in parentheses) are manually provided by the authors to help readers understand the word clusters.

Session The number of sessions in the observation period.

Session$w$Weeks The number of sessions in the last w $(1 \leq w < 8)$ weeks of the observation period.

SessionByDay The number of sessions in each day of week. There are seven different features of this type.

4.2 Response frequency features

Here, we describe the features of the response frequency.

Response(t) The number of system responses with response type t.

FirstResponse(t) **Response(t)** features that are computed by using only the first responses in sessions. Our hypothesis is that first responses in sessions crucially affect user engagement.

LongResponse The number of lengthy responses (more than 50 characters long). Because longer responses require a longer reading time, they are prone to irritate users and consequently decrease user engagement.

ErrorMessage The number of error messages. Voice Assist returns error messages (*Sorry, I don't know.*) when it fails to find appropriate responses to the user's utterances. We consider that these error messages decrease user engagement.

4.3 Time interval features

Here, we describe the features related to the session interval times.

MaxInterval The maximum interval (in days) between adjacent sessions in the observation period.

MinInterval The minimum interval (in days) between adjacent sessions in the observation period.

AvgInterval The average interval (in days) between adjacent sessions in the observation period.

InactivePeriod The time span (in days) from the last utterance to the end of the observation period.

InactivePeriod(t) **InactivePeriod** computed separately for each type t of the last response.

4.4 User profile features

Here, we describe the features of the user's profile information. Since it is not mandotory for users to register their profiles, we expect that those who have provided profile information are likely to be engaged with the system.

Nickname A binary feature representing whether or not the user has provided their nickname.

Birthday A binary feature representing whether or not the user has provided their birthday.

Age Six binary features representing the user's age. They respectively indicate whether the user is less than twenty years, in their 20's, 30's, 40's, or 50's, or is more than 60 years old. Note that these features are available only if the user has provided their birthday.

5 Experiments

In this section, we describe our experimental results and discuss them.

5.1 Experimental settings

We randomly divided the log data into training, development, and test sets with the ratio of 8:1:1. Note that we confirmed that the users in different data sets do not overlap with each other. We trained the model with the training set and optimized hyperparameters with the development set. The test set was used for a final blind test to evaluate the learnt model.

We used the LIBLINEAR tool (Fan et al., 2008) to train the SVM for the dropout prediction and

	Accuracy	F–measure
Baseline	56.8	0.482
Proposed	**77.6**	**0.623**
Utterance frequency	70.2	0.578
Response frequency	54.8	0.489
Time interval	74.6	0.617
User profile	39.9	0.406

Table 6: Classification accuracies and F–measures in the dropout prediction task.

	Precision	Recall
Baseline	0.350	0.774
Proposed	**0.553**	0.714
Utterance frequency	0.458	0.785
Response frequency	0.346	**0.831**
Time interval	0.507	0.789
User profile	0.273	0.793

Table 7: Precisions and Recalls in the dropout prediction task.

the SVR for the engagement level prediction task. We optimized the C parameter on the development set. In the dropout prediction task, we used the `-w` option to weigh the C parameter of each class with the inverse ratio of the number of users in that class. We also used the `-B` option to introduce the bias term.

Next, we describe the evaluation metrics. We used accuracy and F_1–measure in the dropout prediction task. Mean squared error (MSE) and Spearman rank correlation coefficient were used in the engagement level prediction task. These evaluation metrics are commonly used in classification and regression tasks.

We compare the proposed models with baseline method. Because we have no previous work on both tasks, we defined baseline method of our own. The baseline method was trained in the same framework as the proposed methods except that they used only **Session** feature. We chose **Session** for baseline because frequency of use features such as **Session** were shown predictive to similar tasks (Kloft et al., 2014; Sadeque et al., 2015) to prospective user engagement.

5.2 Results

Table 6 illustrates the result of dropout prediction task. The first row compares the proposed method with the baseline. We can see that the proposed

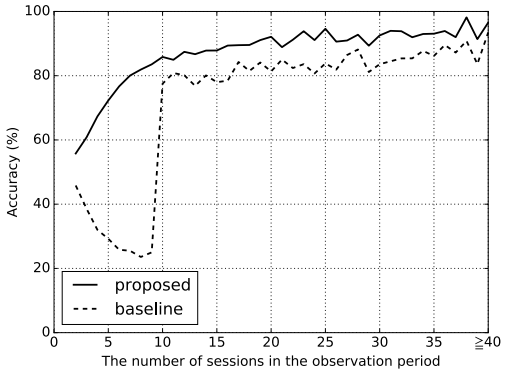

Figure 1: Accuracies per the number of sessions in the observation period of the proposed method and the baseline. The rightmost points represent the accuracy of the users whose number of sessions in the observation period are equal to or more than 40.

	MSE	Spearman
Baseline	0.784	0.595
Proposed	**0.578**	**0.727**
Utterance frequency	0.632	0.693
Response frequency	0.798	0.584
Time interval	0.645	0.692
User profile	1.231	0.146

Table 8: MSE and Spearman's ρ in the engagement level prediction task.

model outperforms the baseline. This indicates the effectiveness of our feature set. The second row illustrates the performances of the proposed method when only one feature type is used. This result suggests that the utterance frequency and time interval features are especially useful, while the combination of all types of features performs the best. We conducted McNemar test (McNemar, 1947) to investigate the significance of these improvements, and confirmed that all improvements are statistically significant ($p < 0.01$).

Table 7 shows the precisions and the recalls of dropout prediction task. As shown in Table 7, the precision of the proposed method performs the best while the recall is worst. We consider that the performance of the precision is more important for our model because taking proactive actions against users who are likely to stop using the system is one of the assumed applications. Taking proactive actions (*e.g.*, push notifications) against users continually using the system might irritate them and de-

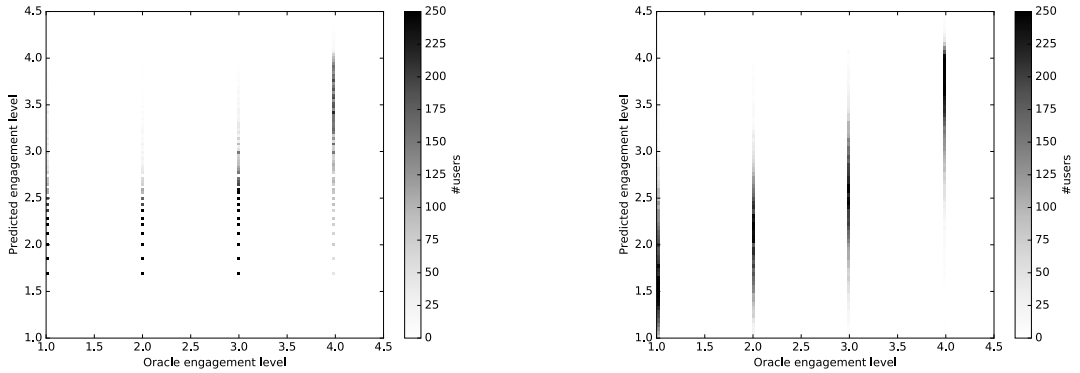

Figure 2: Correlation between the oracle engagement levels and the ones predicted by the baseline method (left) and by the proposed method (right).

crease their user engagement. Therefore, the rate of the users who actually intend to stop using the system in the users predicted as dropout affects the effectiveness of these proactive actions. The result that the precision of the proposed method is 0.553 and that of the baseline is 0.350 is, in other words, using the proposed model improves the effectiveness by 20% absolute in taking these actions.

Figure 1 shows the accuracies per the number of sessions in the observation period of the proposed method and the baseline. The proposed method consistently outperforms the baseline throughout the number of sessions in the observation period. In particular, the proposed method predicts well the dropout of users whose number of sessions is around five compared to the baseline. These results again indicate the effectiveness of the combination of our feature set.

Table 8 shows the result of engagement level prediction task. We again observe similar trends to the dropout prediction task. The proposed method outperforms the baseline. The utterance frequency and time interval features are the most effective, while the combination of all four feature types achieves the best performance in both evaluation metrics.

Figure 2 visualizes the correlation between the oracle engagement levels and the ones predicted by the baseline (left) and by the proposed method (right). We can intuitively reconfirm that the proposed method is able to predict the engagement levels reasonably well.

5.3 Investigation of feature weights

We investigate weights of the features learned by the SVR for figuring out what features contribute to the precise prediction of prospective user engagement.

Table 9 exemplifies features that received large weights for the four feature types. We observe that most features with large positive or negative weights are from the utterance frequency and time interval features. Those include **Session**, **Utterance**, and **InactivePeriod**. It is interesting to see that **UrgedUtterance**, which is based on an utterance type specific to mobile users, also receives a large positive weight.

Further detailed analysis revealed that the proposed model captures some linguistic properties that correlate with the prospective user engagement. For example, **UtteranceTopic36** and **UtteranceTopic23** recieve positive and negative weights, respectively. This follows our intuition since those clusters correspond to greeting and curse words (*c.f.* Table 5). We also observe **Response(WORD CHAIN)**, **Response(QUIZ)** (word association quiz), and **Response(TRIVIA)** (showing some trivia) receive positive weights. This means that playing games or showing some trivia attract users. It is interesting to see that this result is consistent with findings in (Kobayashi et al., 2015). It also follows our intuition that the weight of **ErrorMessage** feature is negative.

5.4 Discussion on length of observation period

Next, we investigate how the length of the observation period affects the prediction performance. We varied the length of the observation periods from one to eight weeks, and evaluated the results (Figure 3).

Figure 3 demonstrates that the model perfor-

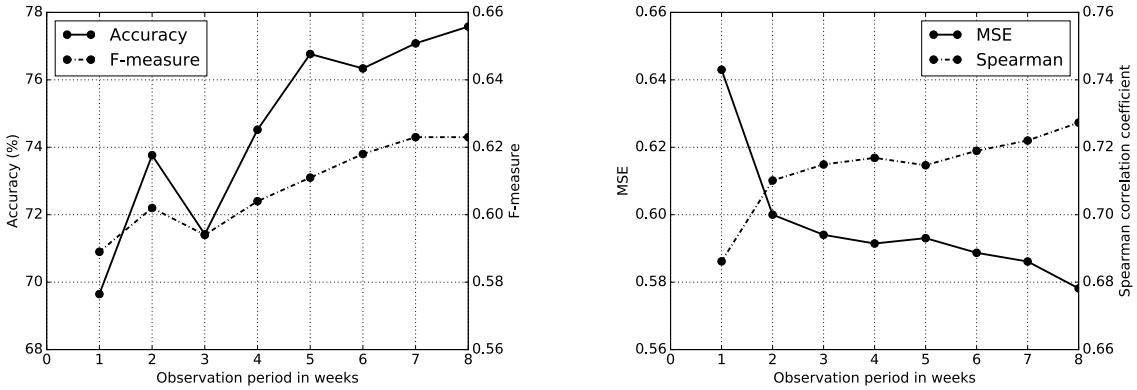

Figure 3: Results of dropout prediction (left) and engagement level prediction (right) across different observation periods (in weeks).

Weight	Feature
0.67	**Session**
0.59	**Utterance**
0.28	**Session7Weeks**
0.26	**UrgedUtterance**
0.02	**UtteranceTopic36**
-0.05	**UtteranceTopic23**
0.08	**Response(WORD CHAIN)**
0.08	**Response(QUIZ)**
0.04	**Response(TRIVIA)**
-0.03	**ErrorMessage**
-0.23	**InactivePeriod(ALARM)**
-0.46	**InactivePeriod**
0.05	**Birthday**
0.04	**Age60s**

Table 9: Feature weights learned by the SVR.

mance generally improves as the observation period becomes longer in both tasks. When we increase the length of the observation period from one week to eight weeks, the accuracy increases by 7.9% in the dropout prediction and Spearman's ρ increases by 4.1 point in the engagement level prediction. The most significant improvements are achieved when we increase the length from one week to two weeks in the three metrics except the F–measure. This suggests that it is generally effective to collect user dialogues of two weeks long, rather than as long as eight weeks or more. This approach would allow to make predictions promptly without waiting for user dialogues to be collected for a long time, while harming accuracy (or other evaluation metrics) as little as possible.

5.5 Application to other intelligent assistants

Here, we discuss how well our approach applies to intelligent assistants other than Voice Assist. The results of this study are considered to apply to other intelligent assistants so long as user logs like the ones in Table 2 are available. The concern is that some attributes in Table 2 may not be available in other systems. In the following, we investigate two attributes, response types and profiles, that are specific to Voice Assist.

We consider that response types like ours are available in user logs of many other intelligent assistants as well. Because our response types mostly correspond to commands issued when operating mobile devices, response types analogous to ours can be obtained by simply logging the commands. Alternatively, it would be possible to employ taggers like (Jiang et al., 2015) to automatically type system responses.

As for profiles, it is likely that similar information is also available in many other intelligent assistants because profile registration is a common function in many IT services including intelligent assistants. For example, Cortana offers greetings and other activities on special days registered by users.[6] Even if user profiles were not at all available, we consider that it would not seriously spoil the significance of this study, because our experiments revealed that user profiles are among the least predictive features.

[6] `http://m.windowscentral.com/articles` (an article posted on Dec. 5, 2015)

6 Related Work

Many dialogue studies have explored the issue of detecting user engagement as well as related affects such as interest and uncertainty (Wang and Hirschberg, 2011; Forbes-Riley et al., 2012; Forbes-Riley and Litman, 2013; Oertel and Salvi, 2013). As discussed in Section 1, these studies typically use a single user utterance to predict whether the user is currently engaged in dialogues with systems. We introduced a new perspective on this line of research by exploring models of predicting prospective user engagement in a large-scale empirical study.

Kobayashi et al. (2015) investigated how games played with intelligent assistants affect prospective user engagement. Although their research interest was prospective user engagement like ours, they exclusively studied the effect of playing game, and left other factors unexplored. In addition, they did not develop any prediction models.

Recently, user satisfaction for intelligent assistants gain attention(Jiang et al., 2015; Kiseleva et al., 2016a; Kiseleva et al., 2016b). Jiang et al. (2015) proposed an automatic method of assessing user satisfaction with intelligent assistants. Kiseleva et al. extended the study of Jiang et al. for prediction (2016a) and detailed understanding (2016b) of user satisfaction with intelligent assistants. Although both satisfaction and engagement are affective states worth considering by intelligent assistants, their research goals were quite different from ours. In their studies, user satisfaction was measured as to whether intelligent assistants can accomplish predefined tasks (*e.g.*, checking the exchange rate between US dollars and Australian dollars). This virtually assesses task-level response accuracy, which is a different notion from user engagement.

Nevertheless, we consider that their studies are closely related to ours and indeed helpful for improving the proposed model. Since user satisfaction is considered to greatly affect prospective user engagement, it might be a good idea to use automatically evaluated satisfaction levels as additional features. The proposed model currently uses **ErrorMessage** feature as an alternative that can be implemented with ease.

Several studies have investigated the chances of predicting continuous participation in SNSs such as MOOC and health care forum (Rosé and Siemens, 2014; Kloft et al., 2014; Ramesh et al., 2014; Sadeque et al., 2015). Unlike those studies, this study exclusively investigates a specific type of dialogue system, namely intelligent assistants, and aims at uncovering usage and/or response patterns that strongly affect prospective user engagement. Consequently, many of the proposed features are specially designed to analyze intelligent assistant users rather than SNS participants.

Our work also relates to the evaluation of dialogue systems. Walker et al. (1997) presented the offline evaluation framework for spoken dialog system (PARADISE). They integrate various evaluation metrics such as dialogue success and dialogue costs into one performance measure function. Although our goal is to predict prospective user engagement and different from theirs, some measures (*e.g.*, the number of utterances) are useful to predict prospective user engagement with intelligent assistants.

7 Conclusion

This paper explored two tasks of predicting prospective user engagement with intelligent assistants: dropout prediction and engagement level prediction. The experiments successfully demonstrated that reasonable performance can be archived in both tasks. Also, we examined how the length of the observation period affects prediction performance, and investigated the trade-off between prediction accuracy and instancy. The future work includes using those prediction models in a real service to take targeted actions to users who are likely to stop using intelligent assistants.

References

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Kate Forbes-Riley and Diane Litman. 2013. When does disengagement correlate with performance in spoken dialog computer tutoring? *International Journal of Artificial Intelligence in Education*, 22(1-2):39–58.

Kate Forbes-Riley, Diane Litman, Heather Friedberg, and Joanna Drummond. 2012. Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 91–102. Association for Computational Linguistics.

Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web*, pages 506–516. International World Wide Web Conferences Steering Committee.

Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Imed Zitouni, Aidan C Crook, and Tasos Anastasakos. 2016a. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 495–505. ACM.

Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016b. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 121–130. ACM.

Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. 2014. Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the 2014 Empirical Methods in Natural Language Processing*, pages 60–65. Association for Computational Linguistics.

Hayato Kobayashi, Kaori Tanio, and Manabu Sassano. 2015. Effects of game on user engagement with spoken dialogue system. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 422–426. Association for Computational Linguistics.

Yujian Li and Bo Liu. 2007. A normalized Levenshtein distance metric. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1091–1095.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability, Vol. 1*, pages 281–297.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Catharine Oertel and Giampiero Salvi. 2013. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pages 99–106. ACM.

Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daumé III, and Lise Getoor. 2014. Learning latent engagement patterns of students in online courses. In *Proceedings of the Twenty-Eighth AAAI Conference Artificial Intelligence*, pages 1272–1278. Association for the Advancement of Artificial Intelligence.

Carolyn Rosé and George Siemens. 2014. Shared task on prediction of dropout over time in massively open online courses. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses*, pages 39–41. Association for Computational Linguistics.

Farig Sadeque, Thamar Solorio, Ted Pedersen, Prasha Shrestha, and Steven Bethard. 2015. Predicting continued participation in online health forums. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 12–20. Association for Computational Linguistics.

Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics.

William Yang Wang and Julia Hirschberg. 2011. Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning. In *Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 152–161. Association for Computational Linguistics.