Evaluation of Machine-led Error Recovery Strategies for Domain Switches in a Spoken Dialog System

Sven Reichel, Ute Ehrlich, André Berton, and Michael Weber

Abstract Spoken dialog systems which include multiple domains or many applications set high requirements for natural language understanding. As the functionality in such systems increase, recognition errors and ambiguous interpretations are likely to occur. However, switching a domain or application by accident reduces user satisfaction and task success rate enormously. Therefore, efficient error recovery strategies need to be applied. In an online study, we evaluated three different machine-led error recovery strategies for in-car infotainment systems. They are varied first in terms of modality (visual and speech) and second in using contextual information. By comparing the strategies, we figured out that asking novice users an open question does not work and they prefer to select the domain from a list of alternatives. This list needs to be minimized in number of items, however, has to contain the requested one. A trade-off between list length and confidence has to be made, based on partial interpreted user utterances and correct predictions of follow-up domains. Furthermore, a choice out of two items requires a graphical visualization, whereby a list performs good with an acoustic presentation and does not need visual elements.

1 Introduction

More and more people are “always on” due to the success of smartphones or other web-enabled devices. The power of these devices increases every year and people use them more than ever. A study by the Nielsen Company shows that app usage in the U.S. rose about 65% from 2012 to 2013 [25]. However, the classical app interaction schema, such as opening an app, interacting with it, and switching to an-

S. Reichel, U. Ehrlich, and A. Berton
Daimler AG, Ulm, Germany
e-mail: sven.reichel@daimler.com, ute.ehrlich@daimler.com, andre.berton@daimler.com

M. Weber
Institute of Media Informatics, Ulm University, Germany
e-mail: michael.weber@uni-ulm.de

other one, is altered by personal assistants (e.g. Apple's Siri[1]or Microsoft Cortana[2]), which are able to recognize and execute user intentions from various domains. For instance, they can search for restaurants, call a selected one to reserve a table, navigate you there, and additionally they will tell you Point-of-Interests on the way, all without switching the app. This is possible because they rely heavily on user-initiated natural speech interaction, which enables users to say whatever they like.

However, "building a dialog management system for the processing of dynamic multi-domain dialogs is difficult" as Lee et al. stated [15]. One crucial point is to identify the domain of interest correctly to process the user's request. This is not an easy task to do, as multi-domain or open-domain Spoken Dialog Systems (SDSs) require large language models, which decrease the speech recognition accuracy and language understanding [5]. Thus, an SDS can never be completely sure, whether the user really intends a domain switch or not.

Processing the domain switch correctly within a multi-domain SDS is crucial to user satisfaction and task success. On the one hand, switching a domain by accident, will require the user to correct or even restart the dialog. On the other hand, not recognizing a domain switch may prevent users from reaching their task goal. While these are more or less user satisfaction issues on a smartphone, for in-car systems they affect the driver's safety seriously. As we have shown in Reichel et al. [19], a non-expected infotainment system behavior results in an increase of driver distraction. As a result, in-car systems need to pay special attention to domain switching and out-of-domain utterances.

Considering this fact, what can in-car systems do if the confidence score of a potential domain switch is low? In this paper, we present different error recovery or clarification strategies, which were evaluated with an online study concerning task success and usability. In Section 2 we provide an overview of existing approaches before presenting our strategies in Section 3. Section 4 describes the study's setup to evaluate our strategies. Results are presented and discussed in Section 5, before we conclude in Section 6.

2 Error Recovery Strategies in Multi-domain SDSs

As Steve Young pointed out in his keynote at SigDial 2014 [26], current SDSs are designed to operate in specific domains, but for accessing web-based information and services, open-domain conversational SDSs are needed. In a previous explorative study [18], we figured out that users do not want to switch between various applications explicitly, instead natural switching between different services should be possible. SmartKom [20] was one of the first SDSs to provide a multimodal interface (Smartakus) for accessing 14 different applications. It is built upon a closed-world ontology and it only understands what is modeled. Recognition errors, or user utterances which are out of domain, are tried to be corrected on a technical level (e.g. query relaxation). Various other technical approaches exist to process domain

[1] https://www.apple.com/ios/siri/, online accessed 2014/09/11

[2] http://www.windowsphone.com/en-us/features-8-1#Cortana, online accessed 2014/09/11

switches and out-of-domain utterances correctly (e.g. [16, 21]). However, they fail for domain ambiguous utterances and even in SDSs using open-world knowledge bases for robust task prediction, situations may occur in which user utterances are ambiguous and an explicit clarification by the user is needed [17].

Bohus et al. [3] analyzed various recovery strategies and identified the "move on" (ignoring the error first and correcting it later on) and "help the user" (providing help messages with sample responses) strategy as good approaches for explicit clarification. However, these are generic clarification strategies, which are used less often in human-human dialogs. Humans prefer context-aware, targeted clarifications to resolve Automatic Speech Recognition (ASR) errors [23]. Skantze's approach [22] also relies heavily on dialog context and partially interpreted user utterances to handle errors in different modules of SDSs. These approaches do not consider domain switches, which often face problems in terms of ambiguities and out-of-domain utterances, thus non-understanding of the complete utterance.

An overview of different error-recovery strategies for multimodal and pervasive systems is provided by Bourguet [4]. She classifies all strategies according to actor, modality and purpose. In our work, the purpose is always to make users clarify the domain their utterance refers to (error correction). Concerning actor and modality, variations of the strategies are developed (see Section 3).

3 Helping the User During a Domain Switch

In a previous experiment [19], we analyzed successful and non-successful domain switches during a driving situation. An error recovery strategy, in which the system takes the initiative and tells users what they can say (Notify and YouCanSay strategy [3]), was compared to them. Concerning task success, this strategy was only 3.3% worse than the successful domain switch, however, it's usability scores and the driver's distraction tended towards the non-successful domain switch. The prompt to tell people what to say was too long and narrative.

Based on these results, three recovery strategies and a reference system were developed to handle uncertainty of domain switches by clarification requests (cf. Appendix):

Reference (REF): An optimal system understands a user's request and executes the desired action. However, as a false domain switch and the requested action would result in severe consequences (e.g. booking a hotel), an explicit confirmation question is always asked. As each participant rates a dialog system on different aspects, we included the reference system to consider these variances. This enables us to compare our strategies with an optimal system.

Ask the User (AU): Asking a user to clarify her intention is always possible for an SDS. Questions can be put in a directed (e.g. "Do you mean Hotel or Facebook") or open-ended (e.g. "Which application are you addressing with your request?") prompt [12]. The AU strategy uses open-ended prompts, which do not restrict users to certain keywords. However, users need to anticipate or know what the system is able to understand (the system's applications).

Domain Choice (DC): Directed dialogs do not require any knowledge of the user as they make clear what the system understands [27]. By having only a limited number of alternatives, the system is able to provide them to the user at once. We propose a choice out of two alternatives. However, in multi-domain SDSs, there might be dialog states in which more than two possible domain switches are likely. This increases the risk to present only wrong alternatives, which slows down the error correction process [24].

Domain List Selection (DLS): If the number of alternatives increases, a list can be presented. While lengthening the prompt, this will reduce the risk to present only wrong alternatives. Users are able to interrupt the prompt by using barge-in after they heard the keyword, which will lead to their task goal. We explained the barge-in and facilitated it by using a short pause after each keyword.

These three dialog strategies enable an SDS to handle cross-domain utterances efficiently. First, they can be used to clarify domain switches in case of low confidence scores. Second, out-of-domain utterances can be classified by the user to the corresponding domain and can be reinterpreted with the right language models.

3.1 Variations of the Dialog Strategies

The success of the dialog strategies may depend on the kind of presentation and use of contextual information. In-car infotainment systems are normally equipped with a display and speakers, so multimodal output can be used. Visual output requires the driver to look at the display and thus increases gaze-based distraction [1, 9]. The REF and AU strategies do not require to present any visual information to the user during a domain switch. However, the most probable follow-up domains in the DC and DLS strategies can be presented using both available modalities. As Suhm et al. showed [24], multimodal error correction strategies are more accurate than unimodal ones. Considering this fact, three different kinds of presentation for each error recovery strategy were developed:

GUI focused (Gui): The idea behind this implementation is to keep the prompts as short as possible. A generic question (e.g. “Say an application name or line number.”) is asked and the alternatives are only displayed on the screen.

Speech focused (S): This variant does not present any dynamic information on screen. Alternatives are only read out as presented in the Appendix.

GUI & Speech (GS): Multimodal output is used to present alternatives on screen and reading them out simultaneously. After selecting an alternative in the DC strategy, it will be highlighted to confirm the selection. The list of alternatives in the DLS strategy is scrolled dynamically, whereby highlighting and reading out is synchronized.

For presenting the alternatives according to the DC and DLS strategy, the system has to decide in which order they appear. As applications are more or less static, it could present them in a fixed order. However, humans usually do not use such a generic clarification strategy and react context-aware [23]. Therefore, we compare two systems, one **with context (withCtx)** and another one **without context (withoutCtx)**. The system with context predicts the most probable follow-up application based on dialog state and user utterance. This application is presented within the two alternatives of the DC strategy and it is added to the top of the DLS list. Without context, the system does present two wrong alternatives for the DC strategy and it inserts the correct application further down of the list, so that scrolling is necessary.

3.2 Hypotheses

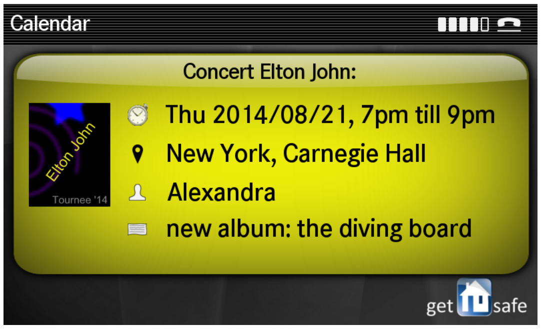
The different variants of our recovery strategies are evaluated concerning usability and task success. It can be assumed that differences exist between strategies, context, and kind of presentation. Table 1 shows the hypotheses. An interesting part is the performance of our error recovery strategies. We assume that significant differences exist between the three strategies and the reference system will perform best (H1). The conditions which consider the context are expected to perform better than the ones without context, as users will reach their task goal more efficiently (H2). Concerning the kind of presentation, no significant differences are assumed in task success because all variants contain the same information. However, users will have preferences for certain kinds of presentation (H3).

Table 1: Hypotheses to evaluate ($=$ no sig. diff.; $\neq$ sig. diff.; $>$ sig. better than)

Hypothesis	Dimension	Task Success	Usability
H1	strategies	$\mathrm{REF} > \mathrm{AU} \neq \mathrm{DC} \neq \mathrm{DLS}$	$\mathrm{REF} > \mathrm{AU} \neq \mathrm{DC} \neq \mathrm{DLS}$
H2.1	context	$\mathrm{DC}_{\mathrm{withCtx}} > \mathrm{DC}_{\mathrm{withoutCtx}}$	$\mathrm{DC}_{\mathrm{withCtx}} > \mathrm{DC}_{\mathrm{withoutCtx}}$
H2.2	context	$\mathrm{DLS}_{\mathrm{withCtx}} > \mathrm{DLS}_{\mathrm{withoutCtx}}$	$\mathrm{DLS}_{\mathrm{withCtx}} > \mathrm{DLS}_{\mathrm{withoutCtx}}$
H3	presentation	$\mathrm{G} = \mathrm{S} = \mathrm{GS}$	$\mathrm{G} \neq \mathrm{S} \neq \mathrm{GS}$

4 Evaluation of Dialog Strategies with an Online Study

The error recovery strategies presented in Section 3 are evaluated with an online user study. This kind of evaluation method allows access to a large number of people in a short time. However, two drawbacks of online studies are missing contextual situation and different interpretation of questions [14]. For now, we neglect the driving situation in favor of many participants and focus on usability as well as task success. In the next step, the best strategies will be implemented and evaluated in a driving simulator. When we designed the GUI, we followed the standardized AAM guidelines [6], which will prevent major driver distraction and prepares the integration into a car's infotainment system. The problem with different interpretations of questions is addressed by using only validated questionnaires.



Concert Elton John:

Thu 2014/08/21, 7pm till 9pm

New York, Carnegie Hall

Alexandra

new album: the diving board

Fig. 1 Dialog context for starting a cross-domain task: participants see a calendar entry showing a concert of Elton John at New York on August 21^{st}, 2014. Alexandra will be there too and a note identifies Elton John's new album. From this dialog state, multiple domain changes are possible (cf. Table 2).

4.1 User Tasks

In a user study, it is crucial to set real tasks for users, as with artificial ones they cannot put themselves into the situation. By using a calendar entry for dialog context (see Figure 1), multiple cross-domain tasks can be imagined. The different semantic values, namely title, date, location, participant, and description can be used to trigger other tasks. Table 2 shows the tasks we used in this study. They are classified into information seeking (inf) and action (act) tasks. This is based on Kellar et al.'s classification schema [13] whereby information exchange and maintenance are grouped together and named action tasks, as they initiate an action.

While tasks occur in real life naturally, in a study users have to be briefed to know their task. This can be achieved through a variety of means. Bernsen and Dybkjaer [2] suggest written instructions or graphically depicted scenarios. However, written instructions prime users to these words and no variances in utterances will be collected. Therefore, we use graphically depicted scenarios.

Table 2: Cross-domain user tasks.

Task	Sem. Value	New Domain	Example User Utterance	Type
T1	Date	Hotel	"Book a hotel for this concert"	act
T2	Date, Location	Weather	"Tell me the weather"	act
T3	Location	Knowledge	"What is the Carnegie Hall?"	inf
T4	Participant	Phone	"Call Alexandra to cancel the appointment"	act
T5	Description	Music	"Play the new album on the Internet radio"	act
T6	Location	Navigation	"Navigate me there"	act
T7	Title	Facebook	"Share this appointment on Facebook"	act
T8	Location	Knowledge	"When was this location established?"	inf
T9	Title	Knowledge	"When was the artist born?"	inf

4.2 Design of the User Study

As described in Section 3, three recovery strategies and one reference system are developed. These are rated and compared with each other by each participant. There are three variations in terms of presentation and two which are affected by context.

By combining the context with each presentation, six variants would emerge. However, varying context in Gui is not reasonable, because of two issues in DLS (cf. Appendix (d)). First, scrolling the list of alternatives would require an additional user interaction step and thus would disadvantage the Gui condition. Second, it is not clear where to present the requested alternative in the list (top, middle, or bottom), because people may start to read at different screen regions. As a result, Gui is implemented in one variant, positioning the requested alternative at different positions. If people are able to compare different variants (Gui, Speech, Gui&Speech), it is likely that they will prefer the multimodal presentation. However, as the system should be implemented within a car, visual distraction is a matter. So we want to figure out whether people need a visual representation or acoustic would be enough. Thus each participant evaluates four dialog strategies in one variant (cf. Table 3).

Table 3: Each participant evaluates one variant

Variant	Presentation	Context	Dialog Strategies
Gui	GUI focused	-	REF, AU, DC, DLS
GS_withoutCtx	GUI & Speech	without	REF, AU, DC, DLS
GS_withCtx	GUI & Speech	with	REF, AU, DC, DLS
S_withoutCtx	Speech focused	without	REF, AU, DC, DLS
S_withCtx	Speech focused	with	REF, AU, DC, DLS

The strategies are evaluated concerning task success and usability. For task success, the user utterances after a system prompt are manually annotated, regarding whether the participant was able to respond correctly or not. Correctly means, an SDS would be able to maintain the dialog flow towards task success. Usability is rated with some questions of the Subjective Assessment of Speech System Interfaces (SASSI) questionnaire [10]. Questions concerning the dimensions Likability, Annoyance, and System Response Accuracy are asked. As participants only rate one system utterance, asking questions concerning the general system performance is not feasible. In addition, three questions from ITU-T Rec. P.851 [11] are asked: help (7.3 Q4), concentration (7.2 Q6), and overall impression. Answers are provided with a 7-point Likert scale from strong disagreement (-3) to strong agreement (+3). The 6 dimensions are averaged to one usability score.

4.3 Procedure of the Experiment

Five variants are required and were implemented with the online tool LimeSurvey[3]. As each participant only takes part in one variant, five groups of participants are needed. However, Hempel et al. [7] observed that users' age, gender and technical experience influences the usability rating and task success of telephone-based SDSs. This means the five groups should have equal populations concerning these attributes. Therefore, we use Hoare et al.'s adaptive random sampling method with

[3] http://www.limesurvey.org, online accessed 2014/09/18

stratification [8] to assign participants to a group after they submitted their age, gender and experience. The link to the study was published via different channels, such as email, mailing lists, personal invitation, flyer, poster, and Facebook.

At the beginning, participants are asked to provide personal data in a questionnaire. After that, the experiment consists of two parts: in the first one participants provide utterances by themselves and in the second one they see videos of sample interactions. Part one requires participants to complete a task with each strategy (strategies are sorted due to learning effects: REF, AU, DC, and DLS) and rate it afterwards. We record the participants' utterances, whereby the system responses are pre-recorded videos (see Appendix for end-to-end sample dialogs). The pre-recorded videos can only be played once, as we want to analyze task success and by repeating the system responses this result would be biased. In addition, barge-in is possible, however, resumption is permitted. After completing the four tasks, participants compare them on a 7-point Likert scale. In the second part of the study, the questionnaires and comparisons are the same as in the first one, but participants judge sample interactions in third person view. This part is randomized, as participants do not need to answer on the questions by themselves, so it does not matter when they see the correct answer for AU in the list of DLS.

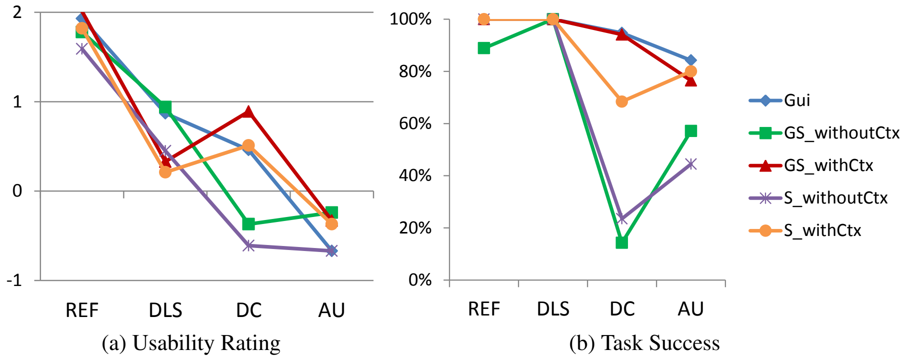
5 Results and Discussion

In the following, evaluation results of the four dialog strategies are shown. We analyzed data from 99 participants (71m/28f), with average age of 30.4 years (*SD*=9.7). They have a medium experience with SDSs (6-Likert Scale, *M*=3.3, *SD*=1.37), but in general they are technical affine (5-Likert Scale, *M*=3.99, *SD*=0.68). 8 participants had problems with their microphone (8m/1f) and 5 aborted after the first part (2m/3f). Nearly all of the tasks were understood correctly by the participants (95%), which confirms our approach with visual task descriptions.

In terms of usability, we assessed four usability scores: (1) rating of each strategy, (2) comparison of the four strategies, (3) rating of each sample interaction, and (4) comparison of the sample interactions. We compared them for each dialog strategy with a repeated measures ANOVA test. No significant differences were found between (1) and (2). However, the AU strategy is rated better in the sample interaction videos than in the interactive part, $F(1,81) = 14.07, p < .001, \eta^2 = .148$ (Helmert Contrast). For DLS this is similar, $F(1,81) = 5.82, p = 0.018, \eta^2 = .067$ (Helmert Contrast). As (1) and (2) are ratings from first person view which are based on real interactions, we use (1) for further comparisons.

5.1 Evaluation of the Dialog Strategies (Hypothesis 1)

The strategies are compared in terms of usability and task success. Figure 2a shows usability scores of the strategies from (1), which differ significantly, $F(3,258) = 113.46, p < .001, \eta^2 = 0.569$. REF is rated best and AU worst. DLS and DC are in between, however DC depends heavily on context (cf. Section 5.2). Concerning task



(a) Usability Rating (b) Task Success

Fig. 2: Results of the interactive part

success (see Figure 2b), REF and DLS are very high and users nearly always reach their goal. In *REF_GS_withoutCtx* some users neglected the explicit confirmation thus task success is lower than in other variants. As with usability, DC depends on context. It can be seen that open questions, such as in AU, do not work properly with novice users. However, AU's usability score of the sample interaction (0.42) shows significant differences compared to the real interaction part (-0.44), $t(85) = -4.74, p < .001$. This leads to the assumption that if users know what application answers the request they will rate the AU strategy better.

5.2 Using Contextual Information (Hypothesis 2)

Analyzing the results concerning contextual information shows importance of context. In DLS and DC the context affects the order of applications, whereby in AU the kind of task is varied (action and information retrieval task). In DLS no significant differences could be identified concerning task success or usability. However, the usability of the without context conditions are rated slightly better than with context. This might be due to the fact that only 37% of participants used barge-in. The others heard the list of applications till the end and had to remember the requested one. As seen in Figure 2, DC depends heavily on context. By showing the requested application, the usability score is on the same level as DLS, otherwise it is worse, $t(70) = 4.25, p < .001$ (GS and S combined). The effect on task success is even worse, only around 20% of the participants reached their goal. In AU there is no significant difference concerning usability. However, task success identified problems in terms of identification of the right application for information retrieval tasks. High variances in the requested application can be seen, such as "Websearch", "Browser", "Wikipedia", or "Google". AU and DC strategy may perform better with expert users, but for novice users their success depends on task type and context.

5.3 Presentation with Different Modalities (Hypothesis 3)

As REF and AU do not require to present any visual information to the user, only DLS and DC are compared. We hypothesized that task success does not depend on the kind of presentation, but usability scores do. As context affects the results (see Section 5.2) we compare conditions with the same contexts (*GS_withCtx* vs. *S_withCtx* and *GS_withoutCtx* vs. *S_withoutCtx*). The usability rating (Figure 2a) shows that in DLS and DC the GUI & Speech variant (GS) is slightly better than the speech focused (S) variant, however, none of these differences are significant (*withCtx*: $t(72) = .86, p = .40$; *withoutCtx*: $t(65) = 1.10, p = .28$). Concerning task success, a difference can be seen between *S_withCtx* and *GS_withCtx* for DC only. If users were asked “Do you want A or B” they often responded “yes”, which cannot be processed by any SDS correctly. A visual representation makes the selection clearer and leads to improvements of task success.

6 Conclusions

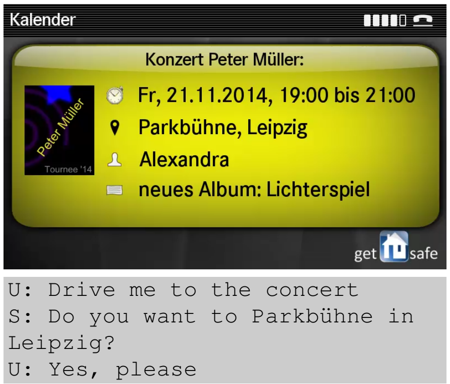
In this work we compared different error recovery strategies for domain switches in SDSs. Obviously, a successful domain switch performs best in terms of usability and task success. However, in case of uncertainty about a domain switch, an SDS should be able to ask the user for clarification. Our results show that an open question, such as “Which application are you addressing with your request?”, does not work for novice users (especially information retrieval tasks are critical). We compared this approach with two recovery strategies in direct prompting style: first a choice out of two alternatives and second, a list selection out of nine items. The results show that the domain choice is a reasonable approach, if the requested application is within the presented alternatives. The domain list allows users to select the right application easily and achieves good usability scores. So far, the dialog strategies are only evaluated with novice users and not in a real driving situation. Expert users, who have learned the interaction schema with machine-led correction strategies, might react appropriately on open questions and thus would be able to interact efficiently. Furthermore, in the car the domain choice might perform better, as duration is a matter. Each second the driver is occupied by the SDS, she might be distracted from the road. In our sample task the domain choice took 6 seconds, whereby the list took 20 seconds. However, domain choice requires a graphical visualization, whereby the list performs good with an acoustic presentation and does not need visual elements.

As a result, each strategy has advantages and disadvantages. Therefore, in the future an adaptive approach has to be considered which adapts the error recovery strategy based on the user (novice or expert) and number of predicted follow-up domains. If only two domains are likely a choice can be used, otherwise a selection list will be better. An intelligent solution has to be developed to limit the number of follow-up domains based on the current dialog state and partial interpreted user utterance. Based on these, an adaptive strategy can be implemented in a car’s infotainment system and can be evaluated in a driving situation.

Acknowledgements The work presented here was funded by GetHomeSafe (EU 7th Framework STREP 288667). We would like to thank the EC for funding the GetHomeSafe project.

Appendix

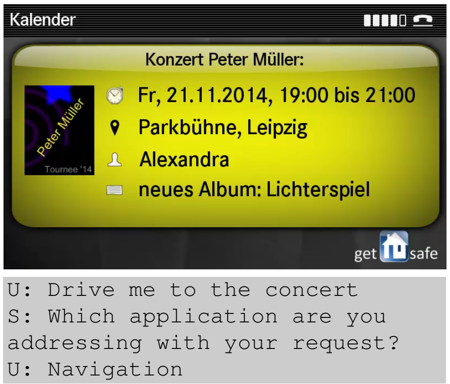
Graphical and speech dialog implementation of the four error recovery dialog strategies (speech dialogs translated from German):



Konzert Peter Müller:
Fr, 21.11.2014, 19:00 bis 21:00
Parkbühne, Leipzig
Alexandra
neues Album: Lichterspiel

```
U: Drive me to the concert
S: Do you want to Parkbühne in
Leipzig?
U: Yes, please
```

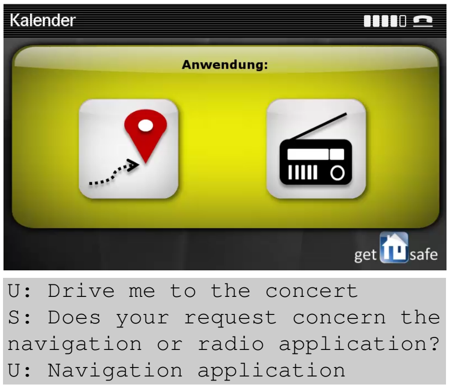
(a) Reference System (REF)



Konzert Peter Müller:
Fr, 21.11.2014, 19:00 bis 21:00
Parkbühne, Leipzig
Alexandra
neues Album: Lichterspiel

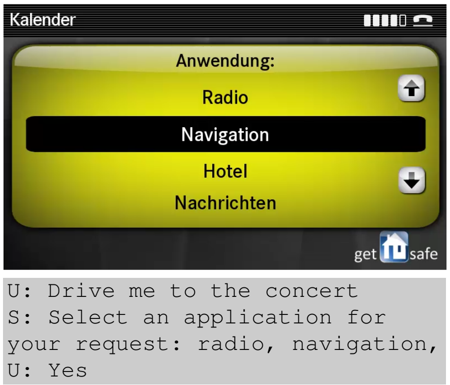
```
U: Drive me to the concert
S: Which application are you
addressing with your request?
U: Navigation
```

(b) Ask the User (AU)



```
U: Drive me to the concert
S: Does your request concern the
navigation or radio application?
U: Navigation application
```

(c) Domain Choice (DC)



```
U: Drive me to the concert
S: Select an application for
your request: radio, navigation,
U: Yes
```

(d) Domain List Selection (DLS)

References

1. Barón, A., Green, P.: Safety and usability of speech interfaces for in-vehicle tasks while driving: A brief literature review. Tech. rep., University of Michigan TRI (2006)

2. Bernsen, N.O., Dybkjaer, L.: Designing Interactive Speech Systems: From First Ideas to User Testing, 1st edn. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1997)
3. Bohus, D., Rudnicky, A.I.: Sorry, i didnt catch that! an investigation of non-understanding errors and recovery strategies. In: Proc. of SIGdial. Lisbon, Portugal (2005)
4. Bourguet, M.L.: Uncertainty and error handling in pervasive computing: A user's perspective. In: Ubiquitous Computing, chap. 3. Babkin, Eduard (2011)
5. Carstensen, K.U., Ebert, C., Ebert, C., Jekat, S., Klabunde, R., Langer, H.: Computerlinguistik und Sprachtechnologie. Spektrum, Akad. Verl. (2010)
6. Driver Focus-Telematics Working Group: Statement of principles, criteria and verification procedures on driver interactions with advanced in-vehicle information and communication systems. alliance of automotive manufacturers (2006)
7. Hempel, T.: Usability of telephone-based speech dialog systems as experienced by user groups of different age and background. In: 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems. Bonn (2006)
8. Hoare, Z., Whitaker, C., Whitaker, R.: Introduction to a generalized method for adaptive randomization in trials. Trials **14**(1), 19 (2013)
9. Hofmann, H., Tobisch, V., Ehrlich, U., Berton, A., Mahr, A.: Comparison of speech-based in-car hmi concepts in a driving simulation study. In: Proc. of IUI. Haifa, Israel (2014)
10. Hone, K.S., Graham, R.: Towards a tool for the subjective assessment of speech system interfaces (sassi). Natural Language Engineering **6**(3&4) (2000)
11. International Telecommunication Union (ITU): Subjective quality evaluation of telephone services based on spoken dialogue systems (2003)
12. Jacko, J.A. (ed.): The human-computer interaction handbook: fundamentals, evolving technologies, and emerging applications, 3. ed. edn. CRC Press, Boca Raton (2012)
13. Kellar, M., Watters, C., Shepherd, M.: A goal-based classification of web information tasks. In: In 69th Ann. Meeting of the American Society for Inf. Science and Technology (2006)
14. Lazar, J., Feng, J.H., Hochheiser, H.: Research Methods in Human-Computer Interaction. John Wiley & Sons Ltd. (2010)
15. Lee, C., Jung, S., Kim, S., Lee, G.G.: Example-based dialog modeling for practical multi-domain dialog system. Speech Communication **51**(5) (2009)
16. Nakano, M., Sato, S., Komatani, K., Matsuyama, K., Funakoshi, K., Okuno, H.G.: A two-stage domain selection framework for extensible multi-domain spoken dialogue systems. In: Proc. of SIGdial. ACL, Stroudsburg, PA, USA (2011)
17. Pappu, A., Rudnicky, A.I.: Predicting tasks in goal-oriented spoken dialog systems using semantic knowledge bases. In: Proc. of SIGdial. Metz, France (2013)
18. Reichel, S., Ehrlich, U., Berton, A., Weber, M.: In-car multi-domain spoken dialogs: A wizard of oz study. In: EACL Workshop Dialog in Motion. Gothenburg, Sweden (2014)
19. Reichel, S., Sohn, J., Ehrlich, U., Berton, A., Weber, M.: Out-of-domain spoken dialogs in the car: A woz study. In: Proc. of SIGdial. Philadelphia, PA, USA (2014)
20. Reithinger, N., Alexandersson, J., Becker, T., Blocher, A., Engel, R., Löckelt, M., Müller, J., Pfleger, N., Poller, P., Streit, M., Tschernomas, V.: Smartkom: Adaptive and flexible multimodal access to multiple applications. In: Multimodal interfaces. New York (2003)
21. Robichaud, J.P., Crook, P.A., Xu, P., Khan, O.Z., Sarikaya, R.: Hypotheses ranking for robust domain classification and tracking in dialogue systems. In: Proc. of INTERSPEECH (2014)
22. Skantze, G.: Error handling in spoken dialogue systems. Ph.D. thesis, KTH Computer Science and Communication (2007)
23. Stoyanchev, S., Liu, A., Hirschberg, J.: Towards natural clarification questions in dialogue systems. In: AISB Symposium on Questions, discourse and dialogue: 20 years after Making it Explicit. London (2014)
24. Suhm, B., Myers, B., Waibel, A.: Multimodal error correction for speech user interfaces. ACM Trans. Comput.-Hum. Interact. **8**(1), 60–98 (2001)
25. The Nielsen Company: Smartphones: So many apps, so much time (2014)
26. Young, S.: Keynote: Statistical approaches to open-domain spoken dialogue systems. In: Proc. of SIGdial. Philadelphia, PA, USA (2014)
27. Zoltan-Ford, E.: How to get people to say and type what computers can understand. Int. Journal of Man-Machine Studies **34** (1991)