Repair Strategies on Trial: Which Error Recovery Do Users Like Best?

Alexander Zgorzelski, Alexander Schmitt, Tobias Heinroth, Wolfgang Minker

Dialogue Systems Group / Institute of Information Technology, University of Ulm
Albert-Einstein-Allee 43, D-89081 Ulm
{alexander.zgorzelski, alexander.schmitt, tobias.heinroth, wolfgang.minker}@uni-ulm.de

Abstract

Extensive research about recovery strategies for misunderstandings and non-understandings within the context of spoken dialogue systems (SDS) has been undertaken in the past and is still going on. Many scientists focus on optimizing the recovery rate using various strategies. It is still not sufficiently explored, how different strategies relate to user satisfaction, and how confused users get with simple strategies such as a reprompt. We carried out an empirical analysis with some of the most promising strategies. In addition to the two common strategies help and reprompt we also evaluated an adapted version of the promising MoveOn strategy. We found that the reactions regarding our different mockup dialogues, especially between computer experts and novices, vary a lot.

Index Terms: spoken dialogue systems, recovery, repair strategy, MoveOn

1. Introduction

Today’s spoken dialogue systems are getting more and more efficient. Automatic Speech Recognizer (ASR) are working very accurate nowadays, yet still far away from the optimum. The recognizers’ progress is usually rated with the two important values Word Error Rate (WER) and recovery rate in case of an understanding problem. In this paper we focus on the latter problem that can be separated into two classes: Non-understandings and misunderstandings.

A non-understanding can relatively easily be detected and solved since it does not lead the SDS into the wrong direction. However, misunderstandings are much more challenging. In case a misunderstanding occurs, theoretically the system still can continue with the dialogue. Thus the challenge is not merely to correct the misunderstanding, but rather the detection of the error. If a user doesn’t know how to signal the misunderstanding or just doesn’t notice the problem, the dialogue may totally go into the wrong direction and may never be concluded successfully. Nevertheless if a user knows how to correct the mistake, there is a good chance that he will succeed. But when he is unsure and tries to continue with a dialogue already going into the wrong direction it is very hard to detect that there was a misunderstanding before and even harder to find the initial momentum that caused the failure.

In this paper, we show the results of our study of the user satisfaction of different recovery strategies in case of a misunderstanding. In Section 2, we talk about the other previous works on this topic. In Section 3, we explain why our motivation for this work the evaluation of user-friendliness is. Section 4 describes our experimental setup followed by our results in Section 5 and corresponding conclusions in Section 6.

2. Related Work

Before we started evaluating recovery strategies it was necessary to conduct a survey in order to find the most technically efficient and intuitively elegant recovery strategies.

Bohus and Rudnicky [2] already gave an accurate overview on this topic. Their goal was to find out how various strategies compare to each other in terms of recovery rate when facing a non-understanding error. Ten different strategies were examined, including the MoveOn strategy explained in the next paragraph. In their work MoveOn was the recovery strategy with the highest recovery rate, followed by a full help strategy, a terse “YouCanSay” and the quite simple reprompt strategy. The second part of that work was an analysis of the efficiency of a recovery policy. Taking a metric into account, which chooses the expected best strategy in every step, the task success rate significantly increased. This shows that there is still some potential for improvements, when adaptively choosing the most promising strategy in a specific situation.

The MoveOn strategy by Skantze [1] is a very interesting approach. In a routing application he tried to ignore a possible uncertainty of a recognized utterance and instead asked a different topic related question. This also led to a great improvement of the recovery rate. But it still has to be clarified whether and how this concept can be adapted to systems for other domains. One reason, for this concept being so effective is that no misunderstanding or non-understanding is signaled. Studies like Swerts [3] showed that once such an error occurred (even if it was successfully resolved), the prosody and pronunciation changed in a way, that the ASR usually has even more problems. So if the user believes in a proper working system, this disturbing factor could be avoided. When it comes to the implementation of a SLDS, it does help a lot to regard that during the design of the dialogue and its recovery strategies.

3. Motivation

The authors mentioned above mainly focused on the efficiency of the various different strategies they observed. However, the question of usability and user friendliness is important in almost the same manner. Bohus and Rudnicky already recognized the most effective strategies. Although that is a crucial point, it shouldn’t be forgotten, that in HCI the user himself must be in the centre of all scientific examinations. If the user gets annoyed, confused or bored because of an inappropriate strategy, this disturbs the dialogue flow and success rate as well. In addition to that chances that the user will use the system next time again will be increased if he wasn’t satisfied with it. A policy like Bohus and Rudnicky tried out is of course a good solution, but they only tried to maximize the success rate and not the user satisfaction. Therefore we propose that it might also be help-
ful to consider the expected user satisfaction when choosing a strategy. For our experiment we have decided to focus on the following three strategies:

**Modified MoveOn** Skantze’s strategy, which is best suited for route guidance and similar applications, can not be adapted as is for Dialogue Systems that collect information. Still the information to be collected is often related. So it might be more effective to skip the confirmations for a couple consecutive questions in order to have more informations when making a combined explicit confirmation. In case this one will be neglected, it could very well be possible to find the utterance based on a combined score, which could include the n-best lists of the ASR, the history of the user and general reasoning to exclude unlikely combinations. In our system the user is asked three questions without the system repeating the users utterance in order to subsequently use an explicit confirmation for that block. An example of a dialogue using this strategy can be found below.

**Reprompt** After every dialogue step the utterance will be immediately implicit confirmed. In case of a misunderstanding the user has to explicitly interfere with signal words such as “wrong” or “false”. This triggers the system to inform the user that the desired correction will be attempted, followed by the simple reprompt of the question.

**Help** This strategy works similar as the one above, except that to the reprompt of the question a help with all in this step necessary information is added.

### 4. Experimental Setup

To find out how these three different strategies perform regarding the user satisfaction a clean and simple setup is necessary. Every factor increasing the complexity of the experiment can bias the results because one strategy might benefit more from that factor than others. Thus we try to minimize the requirements: three well comparable dialogues with the three different strategies, as short and basic as possible. The dialogues are written in VoiceXML 2.1 and the voice platform these dialogues are used with is TellMe.Studio. This forced us to use English as the language of conversation, so we had to consider this for the selection of subjects. The system was called using Skype with a fixed hardware setup. Because the main focus of this study is not the speech recognition itself, we only tested in very good conditions (no noise) with a high-quality headset. To be able to abort the dialogue in case of a problem or never-ending conversation the supervising person was able to listen to the dialogue all the time. However this was only rarely necessary and most of the time the reason for the interruption was quickly found and resolved. In case of a (also very unusual) software or hardware problem, which isn’t the fault of the user or the dialogue system, the affected part of the experiment was repeated.

To compare the three strategies in different scenarios we developed a set of dialogues. After the early decision of comparing the three strategies explained above, we also decided to write three dialogues for each strategy, which resulted in a set of nine dialogues. In order to make the dialogues comparable and, most importantly, not favoring one particular strategy in advance we established a few principles:

- Implement the dialogues as a good engineer would do it
  - Compare every step for best similarity
  - Don’t make one dialogue more likable than others

Still it is almost inevitable that after some cycles the dialogues turn out to be slightly different. So having three dialogues for each strategy is an advantage here because it reduces this variance and results in more accurate and comprehensible results.

The dialogues are on the one hand as rudimental as possible but on the other hand still useful in practice. Each one consists of three questions for the different topics to be solved followed by an overall question to reassure the outcome. The possible utterances for the user’s answers are also kept as short as possible, at best a single word. With such a simple grammar and in good conditions the ASR works very well and makes, if the user behaves suitable, virtually no mistakes. Since we didn’t want to wait for unexpected system mistakes we faked the misunderstandings. This solution worked very well and made it possible to exactly simulate the dialogue situation we wanted to examine. An exemplary dialogue can be seen in Table 1.

In this dialogue everything goes as perfect as we allow it, since we fake the actually correctly recognized ski boots. Our Modified MoveOn dialogue was designed that way, that if everything got understood correctly, the second “guess” will be the right one. If this strategy shall be successful, this should also be the case in real systems most of the time, otherwise the purpose of the strategy hasn’t been achieved.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>Hello, snow rental here! Which equipment do you need?</td>
</tr>
<tr>
<td>User</td>
<td>Ski (ski understood)</td>
</tr>
<tr>
<td>System</td>
<td>What else do you need?</td>
</tr>
<tr>
<td>User</td>
<td>Ski boots (ski boots understood → altered to snowboard boots)</td>
</tr>
<tr>
<td>System</td>
<td>What else do you need or say done to finish the selection?</td>
</tr>
<tr>
<td>User</td>
<td>Poles (poles understood)</td>
</tr>
<tr>
<td>System</td>
<td>What else do you need or say done to finish the selection?</td>
</tr>
<tr>
<td>User</td>
<td>Done</td>
</tr>
<tr>
<td>System</td>
<td>I have ski, snowboard boots and poles! Is that correct?</td>
</tr>
<tr>
<td>User</td>
<td>No</td>
</tr>
<tr>
<td>System</td>
<td>Oh, i have one last guess, else we have to start over. I have ski, ski boots and poles! Is that correct?</td>
</tr>
<tr>
<td>User</td>
<td>Yes, that’s correct.</td>
</tr>
<tr>
<td>System</td>
<td>Great. Have fun on the slope!</td>
</tr>
</tbody>
</table>

For our experiment we found 23 test persons that can be divided into two groups of equal size: one group consisted of the “computer experts” (for example electrical engineers or computer scientists) and the other one consisted of “common people” like teachers, secretaries or older people. Yet we had to make sure in advance their English isn’t a handicap and therefore could influence the results.

Tests of the experimental setup before the real experiment was conducted showed that a number of six dialogues for each subject is ideal. The maximum number of nine dialogues (three domains each combined with the three strategies) would have

1 http://studio.tellme.com
bored and annoyed the volunteer subjects and therefore would have biased the results. We predefined random sequences of six of the nine dialogues, so that every strategy was used one time as the first dialogue, since that one was the one where the participants had most problems with. Later on they got used to the system because of the similar design of the dialogues. To decrease the problems especially in the beginning the users received a short introduction providing all necessary information he would need to successfully conclude the dialogue.

After the completion of one dialogue we wanted the subjects to fill out a questionnaire. We adopted SASSI [4], a standardized method to assess whole dialogues. Because we had a clear imagination of what we want to know from the user, we were able to shorten the comprehensive questionnaire to 17 questions. To keep the subjects as motivated as possible we set up the experiment not to last longer than 30 minutes.

5. Evaluation

To be able to plot clearly represented results we merge three questions into one generalized factor. The following five main quality factors we have examined:

**EFF - Efficiency** How efficient is the system? Is it going quick and smooth or is the system to talkative?

**FRI - Friendliness** How user-friendly is the system? Was it pleasant to use the system or did the user have to concentrate hard?

**REL - Reliability** Is the system reliable or does make many mistakes and has problem understanding the user?

**IRR - Irritating** Has the user been irritated often or did the system work like expected and foreseen?

**MIS - Mistakes** Did the user know what to do in the case of a misunderstanding and was it easy to correct?

Figure 1 shows the overall results of the questionnaire. It can be seen that the users in general preferred the Modified MoveOn system, which leads us a conclusion that it is important to pay attention to not confront the user with strategies he would need to successfully conclude the dialogue.

The good result of our Modified MoveOn has one big catch: The second “guess” was correct in over 90% of the times and sometimes if that one was actually still not correct, the users ignored that and falsely finished the dialogue. One could argue that psychologically it is more frustrating having to do “it all over again” then just correcting the word again. An idea to avoid that could be to ask for the wrongly understood utterance instead of repeating that part of the dialogue. The inhibition threshold of a user doing the dialogue is therefore higher, which leads to wrong dialogue completions. But if the second guess is dependable, the Modified MoveOn strategy bears the advantage, that if an error occurs in the beginning, the user doesn’t notice it and therefore stays positive and his pronunciation natural. The dialogue can be continued more smoothly and chances that the system can resolve itself increase. If this works well, our results show, that this is the system the user are most satisfied with.

We experienced big differences in behavior and aplomb between subjects who feel comfortable with computers and technology and persons who have rare contact with these devices. The experienced users (shown in Figure 2) had way less problems with our dialogues and finished the dialogues most of the time easily with success. If they started to perform well they quickly get less patient and therefore especially appreciate the efficiency of the Modified MoveOn and got annoyed by the help strategy. But if they encounter a bigger problem, they also are happy about a more detailed help, which could always be observed when they desperately looked for help from the supervising author.

Figure 3 shows the more interesting group of test persons. The lesser experienced users reacted more varied. A few had a general disapproval of talking to a machine and therefore had problems with the different dialogues. In addition to that our systems strengthen their attitude since they always make at least one mistake. The greater diversity of the answers of these subjects can very well be seen in the higher standard deviation of most scores in Figure 3 compared to the ones in Figure 2.

Figure 4 compares the two user groups. The question of efficiency is answered quite similar, the experienced user honored the naturalness and the intelligence of the MoveOn system more than the others. It is also interesting to see, that the latter group was a lot more irritated by the surprising guess of the MoveOn system, which leads us a conclusion that it is important to pay attention to not confront the user with strategies he
cannot comprehend. The experts had less problems correcting the utterances in strategies reprompt and help, which probably also is a reason for the big difference of the friendliness scores. Another general observation we made was that women in general were less bothered and more patient by the detailed help prompts than men. Overall we think that especially when dealing with inexperienced people they differ so much, that different user profiles are necessary in order to classify them at best.

6. Discussion

We propose that it is important to pay attention on the users’ characteristics when choosing a recovery strategy. Factors such as background, mood and ability of handling SLDS, are influencing the expected chances of finishing the dialogue smoothly and successfully. Therefore we think a system which also adapts to maximize the users’ satisfaction will therefore increase the task success rate. In general it should be the task of the Dialogue Management to quickly find out who and how skilled the user is in order to do these adaptations. Prosody features and other factors should be comprised to find out if the chosen strategy was the correct one and which ones have to be selected subsequently. For example, if the user seems to be annoyed of the repeating help, it is the task of the system to immediately react to that. In this paper we distinguished different groups of users which prefer and are able to deal with our strategies in different manner. While experts preferred a rather direct and quick procedure, the inexperienced subjects often have more trouble and therefore should be supported accordingly. A policy for recovery strategies like Bohus and Rudicky did is a very promising approach. If it also chooses the strategies with regard on the expected user satisfaction, we think it will work more effectively. Our results can help here, since they show the general tendencies of the different types of users regarding the three analyzed strategies.

7. Acknowledgements

The research leading to these results has received funding from the European Community’s 7th Framework Programme (FP7/2007-2013) under grant agreement n° 216837 and from the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

8. References