Chapter 3: Experiment

3.1
BoRIS and INSPIRE systems

Description:

-BoRIS system: Bochumer Restaurant-Information-System was developed at Ruhr-University Bochum, in Nordrhein-Westfallen, Germany. BoRIS is a mixed-initiative prototype SDS for information on restaurants in the area of Bochum, Germany. The system is able to search for restaurants with respect to five criteria: type of food, location of the restaurant, price range, day of the week and time that the user wants to eat out.

The system architecture follows a structure consisting of a speech recognition component, a speech understanding component, a dialogue manager connected to a restaurant database, and a speech generation component. System components are either available as fully autonomously operating modules, or as wizard simulations providing control over the module characteristics and their performance. The following components are part of BoRIS:

-Two alternatives for speech input: a commercially available automatic speech recognizer (ASR) with key-word-spotting capability, trained to recognize about 395 keywords from the restaurant information domain, including proper names; or a wizard-based recognizer simulation relying on typed input from the wizard. This simulation requires on-line transcriptions of the user utterances to be produced by the wizard, on which controlled "recognition errors" are generated according to a previously measured confusion matrix. The confusion matrix has been determined for the target vocabulary of the recognizer, and it is scaled in order to simulate arbitrary word error rates.

-A rough keyword-matching speech understanding module. It consists of a list of canonical values which are attributed to each word in the vocabulary. On the basis

of the canonical values, the interpretation of the user input in the dialogue context is determined.

-A dialogue manager which is based on a finite-state machine. The dialogue manager either follows an explicit confirmation strategy, or it does not provide any confirmation at all. Help messages explain the values which are permitted in each search category. In the case that restaurants exist which satisfy the requirements set by the user, BoRIS indicates names and addresses of the restaurants in packets of maximally three restaurants at a time. If no matching restaurants exist, BoRIS offers the possibility to modify the request, but provides no specific information as to the reason for the negative response.

-A restaurant database which can be accessed locally as a text file, or through the web via an HTML interface. The database contains around 170 restaurants in Bochum and its surroundings. Searches in this database are based on pattern matching of the canonical values in the attribute-value pairs.

-Different speech generation possibilities: pre-recorded speech files for the fixed system messages, be they naturally produced or with TTS; and naturally-produced speech or TTS for the variable restaurant information utterances. This type of speech generation makes an additional response generation in textual form unnecessary, except for the variable restaurant information and the confirmation parts where a simple template-filling approach is chosen.

The system is accessed by the user through a simulated telephone line with controlled transmission characteristics, using a standard handset telephone. The BoRIS system has been implemented in the Tcl/Tk programming language on the Rapid Application Developer platform provided by the CSLU Toolkit. It is integrated in an auditory test environment at IKA. The environment consists of three rooms: an office room for the test subject, a control room for the experimenter (wizard), and a room for the set-up of the telephone line simulation system. During the experiment, subjects only had access to the office room, so that they would not suspect a wizard being behind the BoRIS system. This procedure is important in order to maintain the illusion of an automatically

working system for the test subject. The office room is treated in order to limit background noise, corresponding to a noise of below 35 dB(A). Reverberation time is between 0.37 and 0.50 s in the frequency range of speech.

Test subjects had to carry out five experimental tasks which were defined via scenarios. The scenarios were used to ensure that the same task goals had to be reached by all test subjects. They were designed in a graphical way to avoid direct priming (e.g. via the chosen vocabulary). The tasks covered the functionality provided by the system, and different situations the system would be used in. Most of the scenarios provide only a part of the information which BoRIS requires in order to search for a restaurant. The remaining requirements have to be selected spontaneously by the test subjects. When no restaurants exist which satisfy the requirements, constraint relaxations are suggested in some scenarios. In this way, it is guaranteed that a task solution exists, provided that the user accepts constraint relaxation. One task is an 'open' scenario where the subjects could define the search criteria on their own, prior to the call. Test subjects were instructed to imagine a realistic usage situation. Nevertheless, most of the subjects will feel to be in a test situation, which differs in several respects from the later usage situation (e.g. by the motivation for using the system, time and money constraints, the fear of being judged and not being the judging subject, etc.). The test situation is reflected in the user factors and the contextual factors of the QoS taxonomy. In order to obtain direct judgments on different quality aspects from the user, a specific questionnaire has been designed (see chapter 8.1). It consists of three parts (A, B and C). Part A collects background information on the subjects (user factors), and on their demands and ideas of the system under test. Part B reflects the spontaneous impressions of the subjects directly after each call. Part C refers to the final impression at the end of the test session, reflecting all experience gained with the system so far.

During each interaction, a log file was produced by the system. This file has been annotated (transcribed and analyzed) by an expert using a specific tool. From the log file and the expert annotation, a large set of interaction variables can be obtained for each dialogue, and the relevance of input variables estimating quality (question 1 on questionnaire - chapter 8.1) can be tested. The variables generated by the annotation tool are described on section 2.2.1. (for more

information, see Möller 2003, Quality of Telephone-based Spoken Dialogue Systems, chapter 6.1.1).

-INSPIRE system: A smart-home-system for spoken-dialogue-based control of domestic devices. (INfotainment management with SPeech Interaction via REmote microphones and telephone interfaces). This system has been developed in the framework of the EU-funded IST project INSPIRE (IST-2001-32746) and has been evaluated by two controlled laboratory experiments carried out at two different test sites, with different groups of test subjects, but addressing the same system. The experiments originally served the evaluation and optimization of the INSPIRE dialogue system, but the following analysis will be limited to the evaluation method itself, and not to the evaluated system.

With the help of the INSPIRE system, a user can operate different devices in the home environment, namely, a TV, a video recorder, an electronic program guide, an answering machine, a fan, three different lights, and the blinds of the test room. The dialogue module consists of a signal pre-processing module (beam forming, echo cancellation, noise reduction), a speech recognizer, a speech understanding module, a dialogue manager, a device interface, and a speech output module.

Speech recognition is usually performed with a commercial recognizer. However, at the time of the reported experiments, the available training (in-sample) data were insufficient, and thus the recognition performance was found to be too low. As a consequence, a human transcriber replaced the speech recognizer in the reported experiments. The speech understanding module matches possible surface forms (keywords) to canonical values (concepts). For dialogue management, generic dialogue nodes have been defined and instantiated according to the piece of information to be gathered, cf. Rajman et al. (2004). The dialogue manager accesses a task model in the form of a database; this database defines the domestic devices and the actions which are under the user's control.

The system versions used in both experiments slightly differed with respect to the animated head: in experiment 1, a film with a real person's head was displayed which moved the lips when the speech output was presented. Because the synchronization between lips and speech was relatively bad, a

simplified character (a puppet sock) was used in experiment 2, for which the synchronization was considered as less critical. In addition, speech prompts and the system vocabulary were slightly adjusted as a result of the first experiment. Apart from these differences, both system versions were identical. Speech prompts were recorded from a male talker and played back with a level of approx. 79 dB(A) at a fixed position in the test room.

The questionnaire to be filled in after each interaction consists of 37 statements which are grouped under 7 categories (overall impression, reaching the desired goals, communication with the system, behaviour of the system, dialogue, personal impression, usability). For the first statement (overall impression of the interaction with the INSPIRE system), a judgment was solicited on a continuous rating scale labelled with five attributes, see chapter 8.2.

3.2
Experimental setup

For all the approaches, a ‘leave-one-out’ method is going to be used in order to train and test the data. For instance, in the BoRIS system, there were 40 users. For all that four approaches that will be used, linear regression, regression and classification trees and neural networks, we train the method with 39 of the users, and then test on the one independent user. Such a procedure allows us to use a high proportion of the available data to train the method. This system is repeated until the last user is evaluated, hence the systematic is repeated 40 times for the BoRIS system. The disadvantage of such an approach is that it requires the training process to be repeated several times (depends on the number of users), which in some circumstances could lead to a requirement for large amounts of processing time (like in the neural networks approach). This is believed to be the best way to evaluate a method, since the training data is maximized.

For the neural networks and the classification trees approaches, a 3-class evaluation system was created, to replace the 0 to 6 evaluation from the questionnaires: ‘bad’ (0 to 2.4), ‘neutral’ (2.5 to 3.5) and ‘good’ (3.6 to 6.0). Table 4 illustrates this 3-class system. This distribution ended up leaving on the BoRIS system 85(43,2%) evaluations as ‘bad’, 66(33,5%) as ‘neutral’ and 46(23,3%) as ‘good’.

Table 4: Example of Target Data, 3-class Question B0 (overall quality)

Dialogue number & evaluation rate (0 to 6)	Bad	Average	Good
1 - 1,5	1	0	0
2 – 3,0	0	1	0
3 – 4,0	0	0	1
4 – 3,0	0	1	0
5 – 1,0	1	0	0
6 – 3,2	0	1	0
7 – 2,5	0	1	0
8 – 1,1	1	0	0
9 – 4,1	0	0	1
10 – 1,0	1	0	0
11 – 3,2	0	1	0
12 – 4,6	0	0	1
13 – 4,0	0	0	1
14 – 2,0	1	0	0