Conducting the Wizard-of-Oz Experiment

Melita Hajdinjak and France Mihelič
University of Ljubljana, Faculty of Electrical Engineering, Slovenia
{melita.hajdinjak,france.mihelic}@fe.uni-lj.si

Keywords: natural-language dialogue systems, Wizard-of-Oz experiment, dialogue-manager evaluation, PARADISE evaluation framework

Received: June 11, 2004

Human-human and human-computer dialogues differ in such an important way that the data from human interaction becomes an unreliable source of information for some important aspects of designing natural-language dialogue systems. Therefore, we began the process of developing a natural-language, weather-information-providing dialogue system by conducting the Wizard-of-Oz *(WOZ) experiment. In WOZ experiments subjects are told to interact with a computer system, though in fact they are not since the system is partly simulated by a human, the wizard. During the development of the weather-information-providing dialogue system this experiment was used twice. While the aim of the first WOZ experiment was, first of all, to gather human-computer data, the aim of the second WOZ experiment was to evaluate the newly-implemented dialogue-manager component. The evaluation was carried out using the PARADISE evaluation framework, which maintains that the system's primary objective is to maximize user satisfaction, and it derives a combined performance metric for a dialogue system as a weighted linear combination of task-success measures and dialogue costs.*

Povzetek: članek opisuje WOZ poskus, tj. testiranje komunikacije človek–računalnik.

1 Introduction

In a nutshell, a dialogue system or a voice interface enables users to interact with some application using spoken language. The application in question, for example, can be a piece of hardware (*command & control systems*) or a kind of database (*interactive voice response*, *information-providing dialogue systems*, *problem-solving dialogue systems*). A detailed overview is given by Krahmer [1]. In this article, we will focus on information-providing, natural-language dialogue systems, which have already been developed for different domains, for instance, restaurant information [2], theatre information [3], train travel information [4, 5], air travel information [6, 7], and weather information [8].

It is generally acknowledged that developing a successful computational model of natural-language dialogues requires extensive analysis of sample dialogues, but the question that arises is whether these sample dialogues should be human dialogues. On the one hand, it has often been argued that human dialogues should be regarded as a guidance and a norm for the design of natural-language dialogue systems, i.e., that a natural dialogue between a person and a computer should resemble a dialogue between humans as much as possible. On the other hand, a computer is not a person. Consequently, human-human and human-computer dialogues differ in such an important way that the data from human interaction becomes an unreliable source of information for some important aspects of designing natural-language dialogue systems, in particular the style and complexity of interaction [9, 10]. This is because the users of natural-language dialogue systems are influenced by the system's language [11], i.e., they often adapt their behaviour to the expected language abilities of the counterpart. Therefore, instead of gathering human-human data, we started the process of designing the Slovenian and Croatian spoken, weather-information-providing dialogue system [12] by conducting the Wizard-of-Oz (WOZ) experiment [10, 13], which is a more accurate predictor of actual human-computer interaction [9]. This is because in WOZ studies subjects are told to interact with a computer system, though in fact they are not. The system is at least partly simulated by a human, the wizard, with the consequence that the subjects can be given more freedom of expression or be constrained in more systematic ways than this is the case in already existing dialogue systems.

During the development of the weather-information-providing dialogue system the WOZ experiment was used twice. While the aim of the first WOZ experiment (section 2) was, first of all, to gather human-computer data, the aim of the second WOZ experiment (section 3) was to evaluate the newly-implemented dialogue-manager component [14]. Consequently, while in the first WOZ experiment dialogue management was still one of the tasks of the wizard, in the second WOZ experiment it was performed by the newly-implemented dialogue-manager component. The differences in the data from both WOZ experiments therefore reflect the dialogue manager's performance. However, this data was evaluated with the PARADISE evaluation framework [15], i.e., a potential gen-

eral methodology for evaluating and comparing the performance of spoken dialogue agents, which maintains that the system's primary objective is to maximize user satisfaction, and it derives a combined performance metric for a dialogue system as a weighted linear combination of task-success measures and dialogue costs.

2 First WOZ Experiment

The aim of the first WOZ experiment [13] was to gather data that would serve as the basis for the construction of the dialogue manager and the speech-understanding component within the developing Slovenian and Croatian spoken dialogue system for weather-information retrieval [12]. However, the first WOZ system consisted of the following components:

- ISDN telephony platform,
- weather-information database,
- wizard's graphical interface [13], designed as an internet application, which included facilities for the playback of predefined spoken responses as well as forms, image fields and handle of some keyboard shortcuts,
- natural-language generation module,
- Slovenian text-to-speech synthesis [16].

Hence, the task of the wizard in the first WOZ experiment was to simulate Slovenian speech understanding (speech recognition and natural-language understanding) and dialogue management. Croatian speech understanding was not performed since only Slovene users were being involved into the experiment. During the experiment, the wizard was sitting behind the graphical interface, listened to users' queries and tried to mediate an appropriate response, which was being successively followed by the natural-language-generation process and the text-to-speech process.

However, a total of 76 Slovene users (38 female, 38 male) were chosen to take part in the first WOZ experiment. The statistical distributions of the users' ages, educations, dialects, the telephone units and the background environments from where the telephone calls were made were chosen to simulate the actual scenarios. The users were given verbal instructions about the general functionality of the system and a sheet of paper containing a description of the tasks they were supposed to complete. They had two scenarios to enact. The first task was to obtain a particular piece of weather-forecast information, such as the temperature in Ljubljana or the weather forecast for Slovenia tomorrow, and the second task was a given situation, such as "You are planning a trip to... What are you interested in?", the aim of which was to stimulate the user to ask context-specific questions. After these two scenarios, users were given the freedom to ask additional questions.

In order to evaluate user satisfaction, users were given the user-satisfaction survey [17] used within the PARADISE framework (section 4), which asks to specify the degree to which one agrees with several questions about the behaviour or the performance of the system (**TTS Performance**, **ASR Performance**, **Task Ease**, **Interaction Pace**, **User Expertise**, **System Response**, **Expected Behaviour**, **Future Use**). The answers to the questions were based on a five-class ranking scale from 1, indicating strong disagreement, to 5, indicating strong agreement. All the mean values are given in table 1. A comprehensive **User Satisfaction** was then computed by summing each question's score, and thus ranged in value from a low of 8 to a high of 40. In the first WOZ experiment, the mean **User Satisfaction** value was 34.08, with a standard deviation of 5.07.

	WOZ1	WOZ2
TTS Performance	4.42	4.29
ASR Performance	4.51	4.29
Task Ease	4.27	3.74
Interaction Pace	3.94	3.76
User Expertise	4.40	4.28
System Response	4.23	3.76
Expected Behaviour	4.31	4.04
Future Use	3.99	3.78
User Satisfaction	34.08	31.96

Table 1: The mean user values in the first (*WOZ1*) and the second (*WOZ2*) WOZ experiment.

The spontaneous speech data, named *Slovenian Spontaneous Speech Queries (SSSQ)*, that was collected during the first WOZ experiment, was transcribed with the Transcriber tool [18]. The transcription was labelled for turns and utterances, and special labels for dialectal words and non-speech sounds were added. An example dialogue is given in table 2.

3 Second WOZ Experiment

The second WOZ experiment was carried out in order to evaluate the performance of the newly-implemented dialogue manager [14], build on the basis of the data collected during the first WOZ experiment. Therefore, all the other components of the system remained the same. Hence, in comparison with the first WOZ experiment, the task of the wizard in the second WOZ experiment was only to simulate Slovenian speech understanding. The wizard was sitting behind the dialogue manager's interface and entered the meaning representation of what the user said.

A total of 68 Slovene users (29 female, 39 male) were chosen to take part in the second WOZ experiment. They were given the same instructions and the same user-satisfaction survey as the users in the first experiment. All the mean user values, which were slightly worse than the

S:	*Hello. The dialogue system for weather-forecast information speaking. Can I help you?*
U:	Yes, I would like the visibility in Štajerska.
S:	*Wait a moment, please ...* [the wizard is choosing her answer] *In Štajerska today - the visibility will be more than 10 km. Is there something else?*
U:	Will there be any snowfalls?
S:	*For which location are you asking for?*
U:	Štajerska.
S:	*Wait a moment, please ...* [the wizard is choosing her answer] *No, in Štajerska today – the sky will clear up. Is there something else?*
U:	What about the weather in Poland in the next few days?
S:	*I have information for today.*
U:	So, how will it be in Poland today?
S:	*Wait a moment, please ...* [the wizard is choosing her answer] *In Varšava, Poland - it is cloudy, the air temperature is -6 degrees Celsius. Is there something else?*
U:	Is there some snow in Poland?
S:	*I do not offer this information. Do you have any other question?*
U:	No, thank you. Goodbye.
S:	*Thank you for your cooperation. Goodbye.*

Table 2: The Slovene-English translation of an example dialogue between a user (U) and the WOZ system (S), recorded during the first WOZ experiment.

values from the first WOZ experiment, are given in table 1. The mean **User Satisfaction** value was this time 31.96, with a standard deviation of 4.99. Note, the difference between the mean **User Satisfaction** values in both experiments is expected since the wizard with her human-level intelligence should had been able to manage the dialogue better than the implemented dialogue-manager component.

The Slovenian spontaneous speech data collected during the second WOZ experiment was named *Slovenian Spontaneous Speech Queries 2 (SSSQ2)*.

In agreement with previous studies [9, 10, 11], we observed that in both experiments the users adapted their behaviour to the expected language abilities of the natural-language-spoken WOZ system. In several dialogues the first question was much longer than the following ones and, in case of repetitions, requested by the system, the speech mode became more articulated, slower and/or louder. Moreover, while the wizard was mediating her response some users made fun of the system, they made comments like "What a voice - terribly", "It is thinking", "It is searching in the computer", and they laugh. But such side remarks certainly would be rather strange in a natural information-providing task because, in both experiments, subjects were basically role playing. They were not real users with real information requirements or real time constraints and telephone bills.

4 Dialogue-Manager Evaluation

The dialogue-manager component [14] was evaluated using the PARADISE framework [15], which maintains that the system's primary objective is to maximize user satisfaction, and it derives a combined performance metric for a dialogue system as a weighted linear combination of *task-success measures* and *dialogue costs* (i.e., *dialogue-efficiency costs* and *dialogue-quality costs*). The PARADISE model of performance posits that a performance function can then be derived by applying multivariate linear regression (MLR) with user satisfaction as the dependent variable and task-success measures, dialogue-efficiency costs, and dialogue-quality costs as the independent variables. Here, user satisfaction, which has been frequently used in the literature as an external indicator of the usability of a dialogue system, is calculated with the survey [17], used in our WOZ experiments.

In order to model the performance of both WOZ systems, we selected 17 regression parameters. First, we computed the task-success measure **Kappa coefficient** (κ) [19], reflecting the wizard's typing errors, and the dialogue-efficiency costs **Mean Elapsed Time** (MET), i.e., the mean elapsed time for the completion of the tasks that occurred within the interaction, and **Number of User Turns** (NUT). Second, the following dialogue-quality costs were selected: **Task Completion** (Comp), i.e., the user's perception of completing the given task; **Mean Words per Turn** (MWT), i.e., the mean number of words per user's turns; **Mean Response Time** (MRT), i.e., the mean system-response time; **Max Response Time** (MaxRT), i.e., the maximum system-response time; **Rejection Ratio** (RR), i.e., the ratio of system moves asking for a repetition of the last utterance; **Help-Message Ratio** (HMR), i.e., the ratio of system help moves; **Check Ratio** (CR) and **Number of Check moves** (NC), i.e., the ratio and the number of system

moves checking some information regarding past dialogue events; **Non-Provided Information Ratio** (NPR), i.e., the ratio of user-initiating moves that do not result in relevant information being provided; **No-Data Ratio** (NDR) and **Number of No-Data Responses** (NNR), i.e., the ratio and the number of system moves stating that the requested information is not available; **Relevant-Data Ratio** (RDR), i.e., the ratio of system moves directing the user to select relevant, available data; **Unsuitable-Initiative Ratio** (UIR), i.e., the ratio of user-initiating moves that are out of context; **Non-Initiating Ratio** (NIR), i.e., the ratio of non-initiating user moves.

When applying PARADISE to the data from the first WOZ experiment to derive a performance equation, we found that **Help-Message Ratio**, **Non-Provided-Information Ratio**, **Task Completion**, **Mean System Response Time**, and **Rejection Ratio** were the parameters that significantly contributed to user satisfaction. On the other hand, the most significant parameters in the second WOZ experiment were **Check Ratio**, **Kappa**, **Mean Elapsed Time**, **Non-Provided-Information Ratio**, and **Task Completion**.

Walker et al. [17] found in their experiments that **Task Completion**, rather than **Kappa**, was a significant factor in predicting user satisfaction, and argued that this was because the user's perceptions of task completion sometimes varied from **Kappa**. In our experiments, **Kappa** only referred to the wizard and **Task Completion** was related only with the first task, which could be the reasons why we did not come to the same conclusion. On the one hand, in these experiments, **Kappa** and **Task Completion** were uncorrelated, but on the other hand, in the second WOZ experiment, **Kappa** was an even more significant predictor of user satisfaction.

However, significant predictors of user satisfaction that did not refer to the wizard were **Help-Message Ratio** and **Non-Provided-Information Ratio** in the first experiment, and **Check Ratio** and **Non-Provided-Information Ratio** in the second experiment. The size of the **Help-Message Ratio** is a consequence of the user's behaviour during the conversation, which is, on the other hand, influenced by the system's level of user-friendliness and cooperation. A user-friendly and cooperative dialogue system should not only play an active role in directing the dialogue flow toward a successful conclusion for the user, it should also be able to take the initiative and to instruct the user if he/she asks for help. However, because some novice users of a dialogue system who are not able to adapt quickly are likely to need instructions provided by the system, **Help-Message Ratio** is expected to reflect user satisfaction. Furthermore, because **Check Ratio** is in a way related to the speech-understanding process, which is usually the most problematic part of a dialogue-system's performance, it is inappropriate to try to decrease it at any price. Consequently, user satisfaction can be remarkably improved only by decreasing **Non-Provided-Information Ratio**. This can be done by preventing the dialogue manager from giving no information before first checking that there is no other available data that might be relevant to the user's request, i.e., the dialogue manager should be as flexible as possible in directing the user to select relevant, available data.

5 Conclusion

In this study we have presented the conducted WOZ experiments, aim of which was to gather human-computer data and to evaluate the dialogue-manager component of the developing, Slovenian and Croatian spoken dialogue system for weather-information retrieval.

The results of applying PARADISE to the data from both WOZ experiments have been given. These have shown that user satisfaction is significantly correlated with the percentage of those user initiatives that did not result in relevant information being provided. We concluded that the ability to direct the user to select relevant, available data is of great importance, and, consequently, that a dialogue system should give no information only if there is no other available data that might be relevant to the user's request.

References

[1] Krahmer, E.J. (2001) *The Science and Art of Voice Interfaces*, Philips research report, Eindhoven, The Netherlands.

[2] Jurafsky, D., Wooters, C., Tajchman, G., Segal, J., Stolcke, A., Fosler, E., and Morgan, N. (1994) The Berkeley Restaurant Project, *Proc. of the 3rd International Conference on Spoken Language Processing*, Acoustical Society of Japan, Yokohama, Japan, pp. 2139–2142.

[3] van der Hoeven, G., Andernach, J., van der Burgt, S., Kruijff, J., Nijholt, A., Schaake, J., and de Jong, F. (1995) A Natural Language Accesible Theatre Information and Booking System, *Proc. of the 1st International Workshop on Applications of Natural Language to Data Bases*, AFCET, Versailles, France, pp. 271–285.

[4] Eckert, W., Kuhn, T., Niemann, H., Rieck, S., Scheuer, A., and Schukat-Talamazzini, E.G. (1993) A Spoken Dialogue System for German Intercity Train Timetable Inquiries, *Proc. of the 3rd European Conference on Speech Communication and Technology*, ISCA, Berlin, Germany, pp. 1871–1874.

[5] Allen, J.F., Schubert, L.K., Ferguson, G., Heeman, P., Hwang, C.-H., Kato, T., Light, M., Martin, N.G., Miller, B.W., Poesio, M., and Traum, D.R. (1995) The TRAINS Project: A Case Study in Building a Conversational Planning Agent, *Journal of Experimental and Theoretical AI*, Taylor and Francis Ltd, pp. **7** 7–48.

[6] Ipšić, I., Mihelič, F., Dobrišek, S., Gros, J., and Pavešić, N. (1999) A Slovenian Spoken Dialogue System for Air Flight Inquires, *Proc. of the 6th European Conference on Speech Communication and Technology*, ISCA, Budapest, Hungary, pp. 2659–2662.

[7] Stallard, D. (2000) Talk'n'Travel: A Conversational System for Air Travel Planning, *Proc. of the 6th Applied Natural Language Processing Conference*, Association for Computational Linguistics, Seattle, USA, pp. 68–75.

[8] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T.J., and Hetherington, L. (2000) Jupiter: A Telephone Based Conversational Interface for Weather Information, *IEEE Transactions on Speech and Audio Processing*, IEEE, pp. **8(1)** 85–96.

[9] Fraser, N.M. and Gilbert, G.N. (1991) Simulating Speech Systems, *Computer, Speech and Language*, Academic Press, pp. **5(1)** 81–99.

[10] Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993) Wizard of Oz Studies: Why and How, *Proc. of the International Workshop on Intelligent User Interfaces*, ACM Press, Orlando, USA, pp. 193–200.

[11] Zoltan-Ford, E. (1991) How to Get People to Say and Type What Computers Can Understand, *Journal of Man-Machine Studies*, Academic Press, pp. **34** 527–547.

[12] Žibert, J., Martinčić-Ipšić, S., Hajdinjak, M., Ipšić, I., and Mihelič, F. (2003) Development of a Bilingual Spoken Dialogue System for Weather Information Retrieval, *Proc. of the 8th European Conference on Speech Communication and Technology*, ISCA, Geneva, Switzerland, pp. 1917–1920.

[13] Hajdinjak, M. and Mihelič, F. (2003) The Wizard of Oz System for Weather Information Retrieval, *Lecture Notes in Artificial Intelligence 2807: Text, Speech and Dialogue*, pp. 400–405. Matoušek, V. and Mautner, P. (eds). Berlin, Springer.

[14] Hajdinjak, M. and Mihelič, F. (2004) Information-Providing Dialogue Management, *Lecture Notes in Artificial Intelligence 3206: Text, Speech and Dialogue*, pp. 595–602. Sojka, P., Kopecek, I. and Pala, K. (eds). Berlin, Springer.

[15] Walker, M.A., Litman, D., Kamm, C.A., and Abella, A. (1997) PARADISE: A General Framework for Evaluating Spoken Dialogue Agents, *Proc. of the 35th Annual Meeting of the Association of Computational Linguistics*, Association for Computational Linguistics, Madrid, Spain, pp. 271–280.

[16] Gros, J., Pavešić, N., and Mihelič, F. (1997) Text-to-Speech Synthesis: a Complete System for the Slovenian Language, *Journal of Computing and Information Technology*, University Computing Centre Zagreb, pp. **5(1)** 11–19.

[17] Walker, M.A., Litman, D.A., Kamm, C.A., and Abella, A. (1998) Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies, *Computer, Speech and Language*, Academic Press, pp. **12(3)** 317–347.

[18] Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001) Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production, *Speech Communication: Special Issue on Speech Annotation and Corpus Tools*, Elsevier Science, pp. **33(1)** 5-22.

[19] Di Eugenio, B. and Glass, M. (2004) The Kappa Statistic: A Second Look, *Computational Linguistics*, The MIT Press, pp. **30**(1) 95–101.