Towards user modelling in conversational dialogue systems: A qualitative study of the dynamics of dialogue parameters

Anna Hjalmarsson

Centre for Speech Technology, KTH Sweden
annah@speech.kth.se

Abstract

This paper presents a qualitative study of data from a 26 subject experimental study within the multi-modal, conversational dialogue system AdApt. Qualitative analysis of data is used to illustrate the dynamic variation of dialogue parameters over time. The analysis will serve as a foundation for research and future data collections in the area of adaptive dialogue systems and user modelling.

1. Introduction

Usability in system design is generally approached by designing for a representative subset of users without considering individual differences [4]. Unfortunately, the stereotyped user that these systems are designed for rarely exists. Our language behaviour is influenced by our individual goals, experiences, skills and surrounding context. We even change style of interaction from one occasion to another depending on contextual circumstances such as stress, purpose and mood. A system with the ability to extract information about its users and that uses this knowledge to adapt its behaviour can be both more efficient and pleasant to use [7]. The motivation for this study is to obtain extended knowledge about the dynamic features of various dialogue parameters in the conversational spoken dialogue system AdApt. This knowledge will serve as a foundation for research and future data collections in the area of adaptive dialogue systems and user modelling.

1.1. Adaptive spoken dialogue systems

The term adaptive systems covers a broad range of interactive systems which adjust to new tasks, situations, users or expressions. Research within this area often focuses on how to support system use in terms of two challenges: error handling and matching expectations with capabilities.

1.1.1. Error handling

A frequently used approach to avoid and recover from errors and misunderstandings in adaptive spoken dialogue systems is to shift dialogue management strategy [9]. If the system can 'sense' error prone dialogues, a shift in dialogue strategy can be used to detect the type of error and to develop a strategy to recover from it.

1.1.2. Matching expectations with capabilities

Another challenging issue in spoken dialogue systems is how to match user expectations with system capabilities. In the ideal system the user knows immediately how to use it. However, this is seldom the case. System experience is one dimension that is relatively often considered in adaptive systems [8] [9]. A user with no system experience will likely need extra guiding. An experienced user, on the other hand, will probably experience such extra guiding as irritating and a waste of time. A system which only gives guidance when needed, that is when lack of guidance leads to errors, can improve the system's performance and the usability. The adaptive dialogue system MIMIC [3] automatically and dynamically shifts dialogue strategy based on the level of user initiative. The system provides more guidance if the user seems insecure and acts passively when the user takes the initiative.

1.2. User modelling

In the field of natural language processing one important objective is to endow a computer with natural language capabilities. In human-human dialogue we acquire and use knowledge about our conversational partners. For machines to interact in the same natural way they too need to acquire information about their conversational partners. Rule-based models use inference rules based on behavioural analyses to make assumptions about the users. These models can sometimes be difficult to apply outside the specific case of study. Statistical models are based on actual samples, observable values of a number of various parameters, to make inferences about an unknown dependent variable. Statistical models require extensive data collections. Walker, Langkilde, Wright, Gorin and Litman [11] predict and identify problematic dialogues with the spoken dialogue system How May I Help You? (HMIHY). The system uses machine learning to generate a model for classifying problematic dialogues early in the interaction based on automatically extracted dialogue features. The model is trained using predefined classes and a number of features that are predicted to determine these classes.

The focus of this paper is on the dynamic variation of various dialogue parameters in relation to the users' subjective ratings of the system and system experience. The organization of this paper is as follows. In section 2 the data collection will be presented along with an overview of the results from a previous empirical evaluation of the system. The parameter set and the foundations for this qualitative study are presented in section 3 which is followed by an analysis of the qualitative results 4 and finally conclusions 5.

2. Dialogue data

The data presented in this study was collected with AdApt, a multi-modal conversational dialogue system developed at CTT (Centre for Speech Technology) at KTH with Telia Sonera as an industrial partner [5].

2.1. Dialogue system and domain

The aim of the AdApt-project is to study human-computer interaction in a multi-modal conversational dialogue system. The system has all the functional components of a spoken dialogue system and resolves anaphora, contextual references and fragmented utterances. The tasks are associated with finding information about available apartments for sale in downtown Stockholm. The system uses speech and mouse clicks as input and a 3-D animated agent [2] that produces lip-synchronized synthetic speech as output. The system also displays the location of the apartments as coloured dots on a map.

User	Hur stor är den gröna lägenheten? How large is the green apartment?
System	Den gröna lägenheten har ett rum. The green apartment has one room.
User	Hur många kvadratmeter? How many square meters?
System	Den gröna lägenheten är på 49 kvadrat. The size of the green apartment is 49 square meters.
User	Vilken våning? Which floor?
System	Den gröna lägenheten ligger på våning 2 The green apartment is on floor 2.
User	Hur många kvadratmeter är den blåa? How many square meters is the blue one?
System	Den blå lägenheten är på 32 kvadrat The size of the blue apartment is 32 square meters.

Table 1: Dialogue example AdApt

2.2. Data collection

The data collection was made from dialogues with 26 subjects. The subjects were all between 20 and 40 years of age. The subjects had no professional experience of speech technology and they had never seen or used AdApt before the experiment. Each user interacted with the system for about 30 minutes. The subjects were instructed to search for information about apartments that they might want to live in, buy or had other interests in. The open task was used to make the interaction between user and system as natural as possible. Task definitions based on the users' dialogue initiatives and labelling of task success were later manually extracted from transcriptions of the dialogues. A similar task definition was used in PROMISE [1]. The data were collected through manual tagging of the dialogues and automatic logging of the interaction. The user's subjective ratings of the system were collected through a questionnaire. A more detailed description of the various parameters and the task definition can be found in Hjalmarsson [6].

2.3. PARADISE evaluation of AdApt

Data were evaluated using PARADISE (PARAdigm for Dialogue System Evaluation) [10]. PARADISE is a general framework for evaluating spoken dialogue systems. PARADISE combines a set of different performance metrics and uses multiple linear regressions to specify how these multiple factors contribute to the overall performance. The model posits user satisfaction as the top-level objective. Predictors of user satisfaction are task success and dialogue costs. Dialogue efficiency and dialogue quality are, in their turn, potential contributors to dialogue costs.

The PARADISE evaluation of AdApt resulted in three statistically significant predictors of user satisfaction: the number of utterances per task, average number of system utterances per user utterance and task success. The interpretations of these results are somehow problematic since the occurrence of depending variables restricted our choice of metrics. Multiple linear regressions presume that the predictor variables are independent of each other. It seams plausible that the quality of the dialogue and the efficiency of the dialogue are somehow related to each other. A correlation matrix showed that some of the parameters were strongly correlated and therefore had to be excluded from the multiple linear regressions to avoid interference.

3. Dynamic evaluation

The PARADISE evaluation resulted in three predictors of user satisfaction. These results were affected by the fact that multiple linear regressions restricted the set of parameters. Moreover, the results do not reveal how the various parameters vary over time. Many of the parameters collected were non-domain specific dialogue features which can be automatically extracted during the interaction. These characteristics make them attractive for user modelling purposes.

3.1. Predicting user satisfaction

For a system to tailor its behaviour to individual users it will be valuable to be able to predict the users' subjective attitudes towards the system early in the interaction. Extended knowledge about the dynamic distribution of various dialogue features is needed to identify what characterizes a dialogue with high as opposed to low user satisfaction. The open task definition restricted us to collect only one user satisfaction at the end of the dialogue. As a result we are not able to model how the users' attitudes vary over the dialogues in the manner prescribed by PARADISE.

3.2. User experience

One purpose of this paper is to study how the users gradually become more accustomed to the system. The subjects in the data collection were all novice users with no prior experience of the system. Moreover, the instructions they received were very few. During the half hour session they become more familiar with the system and gradually acquire a better model of the system domain, the system's vocabulary and its turn-taking features.

3.3. Parameters

For user modelling purposes the first set of metrics were non-domain specific dialogue features which were automatically extracted during the interaction:

- Average number of USER utterances
- Time per USER word
- Number of words per USER utterance
- Average number of SYSTEM utterances

These metrics were all extracted from the system logs and ASR results and no consideration was taken of the transcriptions or manual tagging of the dialogues. Two metrics in the parameter set were based on manual tagging:

- Task success per task
- Average word error rate

Task success is based on transcriptions of the dialogues and system responses. Each system utterance which generated a correct answer that responded to the user's previous request was marked as successful. Task success was used because it was a statistically significant positive predictor of user satisfaction in PARADISE and an indicator of how the users' subjective attitudes vary over time. Task success is also a potential predictor of system experience, i.e. success will likely increase when system experience is gained. Word error rate (WER) is based on the disagreement between manual transcriptions of the dialogues and the system recognitions.

3.4. Method

The dialogues in the AdApt data collection consisted of 6431 spoken utterances of which 2043 (12749 words) were system utterances and 4388 (13837 words) were user utterances. The subjects were divided into two groups based on their subjective ratings of the dialogue. HIGH is the users with the highest user satisfaction ratings and LOW is the users with the lowest ratings. All dialogues were divided into 4 time frames. The length of the dialogues varied slightly and in order to keep the relative first, second, third and fourth part of all dialogues the calculations of the time frames were based on the total length of each specific dialogue. Each time frame is about 7 to 8 minutes long. A non-relative division based on the shortest dialogue (25 minutes) was made just to confirm that the minor variations in length of the dialogues did not affect the results.

4. Results analysis

Qualitative analysis of the data suggests that are some differences in distribution of parameters between dialogues with high user satisfaction and low user satisfaction. The differences appear to be valid over time, which supports the belief that these parameters can be used to predict user satisfaction early in the dialogues.

4.1. Predicting user satisfaction

First, subjects in HIGH tend to have constant higher number of words per utterance (Figure 1). The error bars indicate 95% confidence interval of mean. The metric is a potential predictor of user satisfaction. There is a possibility that utterances with many words is a winning strategy since the AdApt speech understanding modules are designed to handle natural language and not short command like utterances.

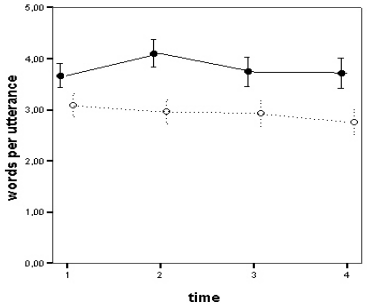

Figure 1: Number of words per utterance

A second potential predictor of user satisfaction is speaking rate. Figure 2 appears to suggest that HIGH speaks slower.

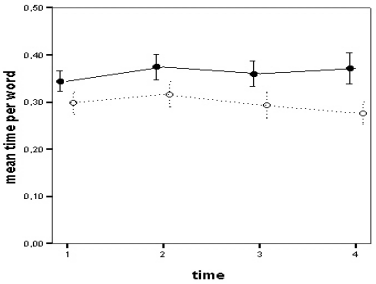

Figure 2: Time per user word

4.2. Predicting system experience

The subjects in the data collection were exposed to a deliberately difficult task. All subjects were novice users, the instructions were few and the task imprecise. The interesting aspect of this set up is that each user was allowed to interact with a system for half an hour, which is a long time. Their attitudes and their conceptual model of the system will likely change a great deal during this period.

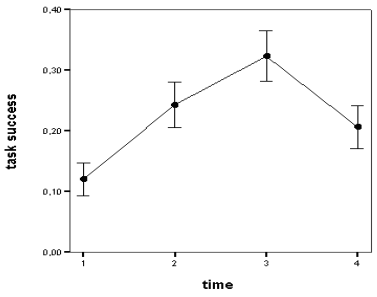

Figure 3: Task success per system utterance

Figure 3 indicates that task success increases during the first 20 minutes. The average number of system utterances is constant over all time frames. The users consequently receive the same number of system responses over all time frames but the amount of correct responses increase drastically during the first 20 minutes. This suggests that the users gradually become more experienced and learn how to extract information from the system. There are several possible interpretations of the decrease in task success at the end of the dialogues. One possibility is that half an hour was too long and that the user loses interest in the system and consequently stops trying. However, this is not supported by the fact that there is only a minor decrease in the number of user utterances at the end. A different interpretation is that the users spend the first part of the dialogue learning one part of the system. When they have figured this part of the system out and are able to use this successfully they might want to explore the system further. This will possibly include trying out new and more complex questions or a new vocabulary that can cause the drop in task success. If this is the case, and the subjects would have been allowed to continue even longer there would likely have been a second peak in task success.

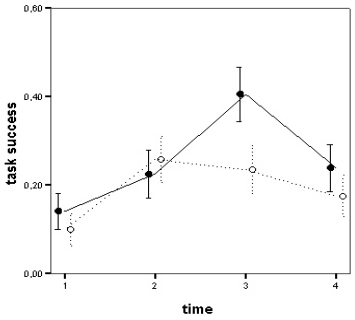

Figure 4: Task success for HIGH and LOW

Figure 4 illustrates that the peak of task success is not as high and appears much earlier in the dialogue for the group with low user satisfaction. The peak for LOW is about 10 minutes into the dialogue while peak for HIGH is 20 minutes into the dialogues. This is an interesting observation. LOW consequently achieves less in terms of task success and appears to stop trying much earlier than HIGH.

Word error rate showed no variation over time and there was only a minor difference between HIGH and LOW. WER was slightly higher at the beginning of the dialogues for LOW. Consequently, WER is not a good predictor of either user satisfaction or system experience in this study. In future data collections a better parameter may be the confidence scores from automatic speech recognition which is an automatically extractable parameter that has been successfully used in other studies [11].

5. Conclusions

This paper has identified potential predictors of user satisfaction. However, the differences between the two groups are rather small and as a basis for adaptation in a conversational dialogue system additional qualitative parameters are likely needed. The variation of task success over time was a rather unexpected but interesting result. To further explore system experience it would be interesting to study what information is actually exchanged in relation to task success.

6. Acknowledgements

This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. The research is also supported by the EU project CHIL (IP506909) and the Graduate School for Language Technology. Many thanks to Rolf Carlson, Jens Edlund and Magnus Nordstrand.

7. References

[1] Beringer, N., Kartal, U., Louka, K., Schiel, F., and Türk, U., "PROMISE - a procedure for multimodal interactive system evaluation". *In LREC 2002 Workshop on Multimodal Resources and Multimodal Systems Evaluation*, Athens, Greece, 2002

[2] Beskow, J., "Talking Heads - Models and Applications for Multimodal Speech Synthesis", *PhD thesis*, 2003

[3] Chu-Carroll, J., "MIMIC: An adaptive mixed initiative spoken dialogue system for information queries", *In 6th ACL Conference on Applied Language Processing,* Seattle, WA, USA, May 2000

[4] Fischer, G., ´´User Modelling in Human-Computer Interaction´´, *User Modelling and User-Adapted Interaction*, 11(1-2):65-68, 2001.

[5] Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., Wirén, M., "AdApt – A Multi-modal Conversational Dialogue System In an Apartment Domain", *In ICSLP 00,* 2000

[6] Hjalmarsson, A., "Evaluating AdApt, a multi-modal conversational dialogue system, using PARADISE", *M.Sc. Thesis*, KTH Royal Institute of Technology, Stockholm, 2002

[7] Kass, R. and Finin T., "Modelling the User in Natural Language Systems" *Computational Linguistics,* 1988.

[8] Komatani, K., Ueno, S., Kawahara, T., Okun, H, G,. "User Modeling in Spoken Dialogue Systems for Flexible Guidance Generation", *In Eurospeech*, pp. 745–748,2003

[9] Van Zanten, G., "User Modelling in Adaptive Dialogue Management", *In: G. Olaszy, G. Németh and K. Erdöhegyi (eds), In the 6th European Conference on Speech Communication and Technology, Eurospeech-99, volume 3, pp. 1183–1186*, 1999

[10] Walker, M., Kamm, C., and Litman, D., "Towards developing general models of usability with PARADISE. Natural Language Engineering", *Special Issue on Best Practice in Spoken Dialogue systems*, September 2000

[11] Walker, M. A., Langkilde, I., Wright, J., Gorin, A. and Litman, D. J., "Learning to Predict Problematic Situations in a Spoken Dialogue System: Experiments with How May I Help You?", *In North American Meeting of the Association of Computational Linguistics,* 2000